1631

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: Scott Arouh)
and Cornelius Diamond)
)
Serial No.: 09/611,220) Group Art Unit: 1631
)
Filed: July 6, 2000) Examiner: Allen. M.
)
For: **NEURAL-NETWORK-BASED IDENTIFICATION, AND APPLICATION, OF GENOMIC INFORMATION PRACTICALLY RELEVANT TO DIVERSE BIOLOGICAL AND SOCIOLOGICAL PROBLEMS, INCLUDING DRUG DOSAGE ESTIMATION**
)
Atty's Docket No.: DIA 0002P)
)

San Diego, California
August 23, 2002

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

Dear Sir:

Transmitted herewith is/are the following document(s) related to the above-identified patent application:

( )	Acknowledgement of Receipt Card	( )	Request for Reconsideration
( )	Disclosure Statement/37 CFR §1.56	( )	Affidavit Under 37 CFR §1.131
( )	Preliminary Amendment	( )	Affidavit Under 37 CFR §1.132
(X)	1 Month Extension of Time Under 37 CFR §1.136 (fee noted below)	( )	Notice of Appeal
		( )	Appeal Brief (in triplicate)
( )	Response Under 37 CFR §1.111	( )	Reply Brief
(X)	Amendment Under 37 CFR §1.115	( )	Certificate of Mailing
( )	Amendment After Final Rejection Under 37 CFR §1.116	( )	Communication
		( )	Change of Address in Application
( )	Power of Attorney by Inventor		

COPY OF PAPERS
ORIGINALLY FILED

CLAIMS AS AMENDED

	Claims Remaining After Amendment		Highest No. Previously Paid For		Present Extra		Rate		Add'l Fee
Total Claims:	4	minus:	34	:	0	x	$ 9	=:	$
Ind. Claims :	2	minus:	12	:	0	x	$42	=:	$
Multiple Dependent Claim Fee								:	$
Fee for Extension of Time (1 Months)								:	$ 55
							TOTAL FEE DUE	:	$ 55

(X) Applicant(s) hereby petition for a 1 month extension of time under 37 CFR §1.136 (fee noted above).
( ) No additional fee is required.
(X) Enclosed is a check for $ 55 in full payment of the above fees. The Commissioner is hereby authorized to charge payment of any additional patent application filing fees under 37 C.F.R. §1.16, 37 C.F.R. §1.17, or patent issue fee under 37 C.F.R. §1.18 associated with this communication or credit any overpayment to Deposit Account No. __-____.

A duplicate copy of this letter is enclosed.

Sincerely yours,

William C. Fuess

William C. Fuess
Registration Number 30,054

William C. Fuess
FUESS & DAVIDENAS
Attorneys at Law
10951 Sorrento Valley Road
Suite II-G
San Diego, California 92121-1613
Telephone: (858) 452-8293
Facsimile: (858) 452-6035
E-mail: fuess@pacbell.net

[X] Attorney of Record
[ ] Filed Under 37 CFR §1.34(a)

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner of Patents and Trademarks, Washington, D.C. 20231, on the date written below.

August 23, 2002	William C. Fuess	William C. Fuess
Date	Typed Name of Person Mailing Correspondence	Signature of Person Mailing Correspondence

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: Scott Arouh)
and Cornelius Diamond)
)
Serial No.: 09/611,220) Group Art Unit: 1631
)
Filed: July 6, 2000) Examiner: Allen. M.
)
For: **NEURAL-NETWORK-BASED IDENTIFICATION, AND APPLICATION, OF GENOMIC INFORMATION PRACTICALLY RELEVANT TO DIVERSE BIOLOGICAL AND SOCIOLOGICAL PROBLEMS, INCLUDING DRUG DOSAGE ESTIMATION**
)
Atty's Docket No.: DIA 0002P)
)

#5/13
Plunkett
9/15/02

San Diego, California
August 23, 2002

AMENDMENT UNDER 37 C.F.R. §115

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

Dear Sir:

Timely in response to the first Office Action mailed April 23, 2002, the time for response to which is extended by the accompanying Petition, please amend the above-identified patent application as follows:

N.G. THERE IS NO MARKED-UP COPY

In the Specification

Please replace pages 61 and 62 of the application as filed with the attached two pages 61 and 62.

In The Claims

Please amend claims ~~10~~, 14 and 15.

REMARKS

Claims 9, 10 and 14-15 are in the application.

Reconsideration and reexamination are respectfully requested.

1. Requirement for Restriction Under 35 U.S.C. §121

Per the previous Office Action mailed October 1, 2001, and Applicants' response thereto mailed November 1, 2001, a Requirement for Restriction Under 35 U.S.C. §121 remains in effect.

Applicants having affirmed their election of invention IV of claims 9-10 and 14-15, claims 1-8, 11-13 and 16-26 remain canceled without prejudice as directed to an un-elected invention.

2. Objections to the Specification

The specification is objected to at pages 61-62.

There is an unwarranted page break, and a slight break in the text, but no material is missing. Replacement pages 61 and 62 are attached. No substantive new material is added.

3. Rejections Under 35 U.S.C. §112, First Paragraph

Claim 10 was rejected under 35 U.S.C. §112, first paragraph.

The Examiner believes "that the information required to practice the [claimed method of the] invention is not available, and does not exist" (Office Action page 3, lines 8-9).

In fact, (i) such information does exist, and (ii) is well known to practitioners of the art to which the invention pertains. Further, (iii) equivalent information regarding "genomic data including alleles and/or characteristic SNP patterns" to that which is suitable for use in the present invention **has** before the filing date of Applicants application been used to realize "Pharmogenetic prediction of clozapine response" (see section 3.3 and EXHIBIT C hereinafter). Still furthermore, (iv) using publicly available data regarding "genomic data including alleles and/or characteristic SNP

patterns", Applicants have reduced their own invention to operative practice.

3.1 Applicants Have Reduced their Claimed Invention to Operative Practice

Considering these points roughly in reverse order (and arguably also in reverse order of significance to this Applicants' argument of their compliance with 35 U.S.C. §112, first paragraph, and of the patentability of their claimed invention), Applicants, presently doing business as Prediction Sciences, Inc. (see <http://www.predictionsciences.com> attach as EXHIBIT A a copy of their own final report in the matter of an SBIR/STTR contract of the U.S. Government operating through the National Science Foundation.

This January 18, 2002, report is, or course, **not** prior art to Applicants' invention. It is introduced for what it shows and describes about 35 U.S.C. §112, first paragraph, issues of concern to the Examiner.

The "Hypertension Response Prediction" described at page 4, paragraph 1, of the report as using "neural nets... to predict response or non-response of the patient to anti-hypertensive therapy" is **clearly** the "method of identifying from the genomic data of an individual organism an adverse reaction to a therapy for at least one disease of the organism" which is set forth in the preamble to claim 9. The report also relates to the "method of predicting an optimal drug dosage and/or drug efficacy for a particular individual patient in respect of genomic data, including alleles and/or characteristic SNP patterns" of Applicants' claim 10.

Importantly to the Examiner's assertion that "information required to practice the [claimed method of Applicants'] invention is not available, and does not exist" (Office Action page 3, lines 8-9), the report at page 13, paragraph 1 explains

the "publicly available data set" -- the AML/ALL data of Golub, et al. available at the time Applicants' application was filed -- used by Applicants in the training of a neural network in accordance with their method. This information of Galub, et al. regards in particular the gene expression of leukemia. It thus differs slightly from the SNP's Applicants prefer for the training of their neural network. The Galub, et al., information **is**, however, a type "genetic information" claimed by Applicants. (Ergo, this information supports Applicants' invention **as claimed**. Moreover, as to Applicants' claimed "genetic information" being broader than simply alleles and/or characteristic SNP polymorphisms, see section 4 hereinafter.)

Note also at page 30, paragraph 2, that "samples from... [the] genomic DNA patient data base [newly developed by the data-source-supplying branch of Pharsight corporation by Applicants as Pharsight's first customer for this data was being developed concurrently with Applicants reported investigations]". This means, or course, that this data -- unlike the AML/ALL data of Golub, et al. discussed above -- was **not** yet available as of the July 6, 2000, filing date of Applicants' application.

Such data **can** be obtained, however, by means both routine and known to practitioners of the art to which Applicants' invention pertains. Note, for example, at the terminal paragraph of page 10 of Applicants' report the dearth of "genome-side allelic association studies", making reference to many reported investigators and investigations that are in process of developing the same -- at least on a piecemeal basis (which is all Applicants' claimed method requires to work).

Although the Examiner is **not** understood to be saying the realization of neural networks is inadequately disclosed under 35 U.S.C. §102, first paragraph, nor beyond the ability of a practitioner of the art to which the invention pertains to realize without undue experimentation, **if** there is any question

in this regard, then please note that the EXHIBIT A report discusses at Appendix B that the training of neural networks in accordance with Applicants invention may be realized by, inter alia, existing Matlab software.

3.3 The Reference Prior Art of Aranz, et al.

Although the attached EXHIBIT B prior art of Aranz, et al. is **not** -- for making no teaching nor any suggestion of the use of a neural network -- particularly relevant to the patentability of Applicant's invention, this prior art, **taken alone and without more**, negates the Examiners's assertion that Applicants' specification does not teach sources of genomic data sufficient to permit practice of Applicants' invention by a practitioner of the art to which the invention pertains.

Namely, the "combination of six polymorphisms" set forth in the preamble were derived from some nineteen such polymorphisms listed in Table 2, and which were considered. Therefore, and without benefit of a neural network optimization, Aranz, et al., are using selected **genomic data**.

And the purpose to which Aranz, et al. are putting this data is, of course per the title of their paper, the "pharmacogenetic prediction of clozapine response". This paper taken alone shows that is was within the ability (as of one month prior to the filing of Applicants' application) for investigators in pharmacogenetics to both (i) had to hand patient genetic **and** pharmacologic response data, and (ii) to associate from this data which polymorphisms were relevant to patient responses to drug therapy (i.e., clozapine), and vice versa (i.e., which polymorphisms usefully predict patient responses to clozapine).

3.4 Still Further Concurrent Developments Show That the Source Genomic Data Upon Which Applicants' Claimed Invention is Suitably Operative Can be Straightforwardly Obtained (Albeit

Possibly with Some Economic Cost)

Still further developments since Applicants' application was filed reveal that activities in progress prior to Applicants' filing date were making use of same or like genomic data to that upon which Applicants' claimed invention is suitably operative, and that this information can be straightforwardly obtained -- albeit possibly at some economic cost.

A current year press release of Genaissance Pharamaceuticals, Inc. is attached as EXHIBIT C. Also included in this EXHIBIT C is a commentary by Applicants on this press release where Applicants find that, inter alia, Genaissance Pharamaceuticals, Inc. has found "29 haplotypes spread over 27 genes", including "25 with drug-specific significance". There is thus offered indirect evidence that patient data regarding both genomics and pharmacogenetic response(s) has been available/is routinely being developed.

Finally, the EXHIBIT D news article -- referencing papers **prior** to Applicants' filing date -- likewise shows that data in respect of **both** genomics and (associated) patient response to therapies, etc. **is** and has been available.

3.5 Applicants' Invention

Applicants do **not** claim to have been the first to recognize the existence of some association, and/or any actual associations, between, on the one hand, genomic data, particularly including "alleles and/or characteristic SNP patterns" and, on the other hand, "an adverse reaction to a therapy for at least one disease" (claim 9) and/or "optimal drug dosage and/or drug efficacy for a particular individual patient" (claim 10).

Applicants invention is clearly

> "**training a neural network** on numerous examples of (i) genomic data including alleles and/or characteristic SNP

patterns, and corresponding (ii) historical drug dosage results including optimal drug dosages, for a multiplicity of patients so as to make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data, including alleles and/or characteristic SNP patterns, to (ii) drug dosage results including optimal drug dosages; and

exercising the trained neural network on the genomic data, including the alleles and/or characteristic SNP patterns, of a particular individual patient to predict an optimal drug dosage for the particular individual patient from among the optimal drug dosages to which the neural network was trained." (claim 10)

In simplest terms, Applicants are the first to apply neural networks to the diverse purposes of genomic analysis.

4. Rejections Under 35 U.S.C. §112, Second Paragraph

Claims 10 and 14-15 were rejected under 35 U.S.C. §112, second paragraph.

Claims 14 and 15, found indefinite for depending upon canceled claims, are amended.

The preamble of claim 10 is revised to state "**therapeutically** optimal drug dosage". (Note that, with the extreme flexibility offered by Applicants' trained neural network, drug dosage **could** be optimized on something else, such as avoidance of toxic side effects!)

The Examiner is correct in finding that the "data input for drug dosages" within the body of claim 10 can include (as claimed) but **does not require** data for optimal drug dosage (as per the preamble of Applicant's claim 10). Applicants' claimed method can, from the data for many patients, derive an optimal drug dosage -- at least in respect of patient genomics -- **even though** the optimal drug dosage for this **or for any** patient may

not previously have been known (i.e., may not be known **even without** consideration of patient's(s') genetics). Of course, it does not hurt the efficacy and effectiveness of Applicants' claimed methods when, and if, the optimum drug dosage for patients of a sample **is** known. Note that in many case it can be said that, at least in the professional judgment of the attending and drug-prescribing physician, the optimal drug dosage is always "known" at least by the lights of this physician.

The phrase "genomic data, including alleles and/or characteristic SNP patterns" is found indefinite. "[A]lleles and/or characteristic SNP patterns" are but examples of genomic data. The phrase can be deleted by amendment if required, leaving only "genomic data". This genomic data can include, for example, the HLA haplotypes discussed in the EXHIBIT D news release.

The Examiner is correct in observing that the claimed "drug dosage results" are **not** required to be associated with (1) alleles, nor (2) characteristic SNP patterns. The "drug dosage results" **could** be associated with other genomic data such as, by way of example, the aforementioned HLA haplotypes. This does **not** mean, however, that "the allele could be concerned with hair color and drug dosages concerned with optimal aspirin dosage for headache relief" -- or at least that **all** alleles (**if** they are the "genomic information") could be so concerned. In actual fact, alleles within a database used in and by Applicants' method for, by way of example, determining "optimal aspirin dosage for headache" **could** indeed be concerned with hair color! Applicants' neural-network-based method "sorts out" the significant alleles (and/or other genetic information) from the insignificant. **If** alleles determining hair color prove **un**important to consumption of aspirin than the neural network will serve to filter the same from consideration as to determination of optimal dosage. However, who knows? Hair color, being strongly correlated with

both race and ethnicity, may actually show a useful correlation with aspirin consumption!

This comment having been made, Applicants do, however, take the Examiner's "point" in his finding that the word "corresponding" of Applicants' claim 10 is either (i) too loose, and/or (ii) modifying of the wrong phrase, in stating a relationship between "(i) genomic data including alleles and/or characteristic SNP patterns, and... (ii) historical drug dosage results [including optimal drug dosages}" (claim 10).

Applicants accordingly amend their claim to state: "which historical drug dosage results are related to **at least some of** the genomic data so as to make a trained neural network" (boldface added).

Note again that, in accordance with Applicants' claimed invention, exactly what genomic data is corresponding to, or related to, or correlatable with, exactly what genomic data (or alleles, or characteristic SNP patterns) is not, and **need not be** precisely known, **nor even guessed at**, prior to entrance into Applicants claimed method. This is the power of training a neural network! The neural network training will come to embody and to weigh heavily in the trained neural network those associations are important, and will minimize or obviate consideration of associations that are **un**important.

It is thus important, and distinctly **non**-trivial, that Applicants' method does "make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data, including alleles and/or characteristic SNP patterns, to (ii) drug dosage results including optimal drug dosages". Indeed, this step alone is "half the battle"! This claimed process is distinctly **un**like the mere "choices" of Aranz, et al.

Applicants attempt to succinctly and precisely claim this attribute of their invention, which attribute is neither taught nor suggested by the art of reference taken in any combination.

5. Summary

The present amendment and remarks have overcome and is discussed each of the bases for the rejections presented in the Office Action. No new subject matter has been introduced by the present amendment.

In consideration of the preceding amendment and accompanying remarks, the present application is deemed in condition for allowance. The timely action of the Examiner to that end is earnestly solicited.

Applicant's undersigned attorney is at the Examiner's disposal should the Examiner wish to discuss any matter which might expedite prosecution of this case.

Sincerely yours,

William C. Fuess
Registration Number 30,054

FUESS & DAVIDENAS
Attorneys-at-Law
10951 Sorrento Valley Road
Suite II-G
San Diego, California 92121-1613
Telephone: (858) 452-8293
Facsimile: (858) 452-6035
E-mail: fuess@pacbell.net

William C. Fuess
[X] Attorney of Record
[ ] Filed Under 37 CFR §1.34(a)

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner of Patents and Trademarks, Washington, D.C. 20231, on the date written below.

August 23, 2002	William C. Fuess	
Date	Typed Name of Person Mailing Correspondence	Signature of Person Mailing Correspondence

CLAIMS (IN AMENDED FORM)

What is claimed is:

9. A method of identifying from the genomic data of an individual organism an adverse reaction to a therapy for at least one disease of the organism,

the method particularly serving to identify a relationship between, on the one hand, (i) any adverse reaction to at least one therapy for at least one disease of an organism, and, on the other hand, genomic data of the organism in the form of two or more alleles and/or SNP pattern(s) of the organism,

the method still more particularly serving to determine which of a large number of alleles as variously occur in the genomic data of a large number of individual organisms are, in actual fact, relevant, both individually and in combination, to certain biological and social variables of these organisms, including the adverse reaction to the at least one therapy for the at least one disease of these organisms,

the method comprising:

1) constructing a neural network suitable to map (i) genomic data of individual organisms as inputs to (ii) historical incidences of responses, including adverse reactions, to therapies for diseases of the individual organisms as outputs;

2) training the constructed neural network on numerous examples of (i) genomic data, as corresponds to (ii) historical incidences of responses including adverse reactions to therapies for the diseases of a multiplicity of individual organisms, so as to make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data to (ii) incidences of therapeutic responses, including adverse reactions, to therapies for the diseases of the organisms; and

3) exercising the trained constructed neural network in respect of a particular therapy for a particular disease of a particular organism, from among the therapies and the diseases to which the neural network was trained for organism including the particular

organism, in order to identify any relationship between (i) any adverse reaction among the responses to the particular therapy, and (ii) genomic makeup of the particular organism;

wherein the neural network is constructed for, and trained on, more organisms than the individual organism on which it is exercised.

10. (Amended) A method of predicting a[n] <u>therapeutically</u> optimal drug dosage and/or drug efficacy for a particular individual patient in respect of genomic data, including alleles and/or characteristic SNP patterns, of the particular individual patient, the method comprising:

training a neural network on numerous examples of (i) genomic data including alleles and/or characteristic SNP patterns, and [corresponding] (ii) historical drug dosage results, including optimal drug dosages, for a multiplicity of patients <u>which historical drug dosage results are related to at least some of the genomic data</u> so as to make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data, including alleles and/or characteristic SNP patterns, to (ii) drug dosage results including optimal drug dosages; and

exercising the trained neural network on the genomic data, including the alleles and/or characteristic SNP patterns, of a particular individual patient to predict an optimal drug dosage for the particular individual patient from among the optimal drug dosages to which the neural network was trained.

14. (Amended) The method according to claims 9[,] <u>or</u> 10[, 11, 12, or 13]

wherein the training is automated by computerized programmed operations using a genetic algorithm.

15. (Amended) The method according to claims 9[,] <u>or</u> 10[, 11, 12, or 13] wherein the training is automated by computerized programmed operations using a genetic algorithm reduced in computational complexity by including the steps of:

grouping alleles and/or characteristic SNP patterns into families as are defined by (i) having similar expression patterns, or (ii) being turned on and off by another gene, or (iii) both having similar expression patterns and being turned on and off by the same gene; and

starting training of the neural network with the genetic algorithm by using the families so created as single inputs to the neural network, the training with the genetic algorithm continuing repetitively until, families of greater and lesser significance being identified, it becomes computationally possible to train the neural network to genomic data consisting of individual alleles and/or characteristic SNP patterns;

wherein partitioning of all alleles and/or characteristic SNP patterns into families permits training of the neural network in a hierarchy of stages, first to the families and only then to the individual alleles and/or characteristic SNP patterns.

CLAIMS (IN PLAIN TEXT FORM)

What is claimed is:

B1

9. A method of identifying from the genomic data of an individual organism an adverse reaction to a therapy for at least one disease of the organism,

the method particularly serving to identify a relationship between, on the one hand, (i) any adverse reaction to at least one therapy for at least one disease of an organism, and, on the other hand, genomic data of the organism in the form of two or more alleles and/or SNP pattern(s) of the organism,

the method still more particularly serving to determine which of a large number of alleles as variously occur in the genomic data of a large number of individual organisms are, in actual fact, relevant, both individually and in combination, to certain biological and social variables of these organisms, including the adverse reaction to the at least one therapy for the at least one disease of these organisms,

the method comprising:

1) constructing a neural network suitable to map (i) genomic data of individual organisms as inputs to (ii) historical incidences of responses, including adverse reactions, to therapies for diseases of the individual organisms as outputs;

2) training the constructed neural network on numerous examples of (i) genomic data, as corresponds to (ii) historical incidences of responses including adverse reactions to therapies for the diseases of a multiplicity of individual organisms, so as to make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data to (ii) incidences of therapeutic responses, including adverse reactions, to therapies for the diseases of the organisms; and

3) exercising the trained constructed neural network in respect of a particular therapy for a particular disease of a particular organism, from among the therapies and the diseases to which the neural network was trained for organism including the particular

organism, in order to identify any relationship between (i) any adverse reaction among the responses to the particular therapy, and (ii) genomic makeup of the particular organism;

wherein the neural network is constructed for, and trained on, more organisms than the individual organism on which it is exercised.

10. (Amended) A method of predicting a therapeutically optimal drug dosage and/or drug efficacy for a particular individual patient in respect of genomic data, including alleles and/or characteristic SNP patterns, of the particular individual patient, the method comprising:

training a neural network on numerous examples of (i) genomic data including alleles and/or characteristic SNP patterns, and (ii) historical drug dosage results, including optimal drug dosages, for a multiplicity of patients which historical drug dosage results are related to at least some of the genomic data so as to make a trained neural network that is fit, and that possesses a measure of goodness, to map (i) genomic data, including alleles and/or characteristic SNP patterns, to (ii) drug dosage results including optimal drug dosages; and

exercising the trained neural network on the genomic data, including the alleles and/or characteristic SNP patterns, of a particular individual patient to predict an optimal drug dosage for the particular individual patient from among the optimal drug dosages to which the neural network was trained.

B1 Cont'd

B2

14. (Amended) The method according to claims 9 or 10

wherein the training is automated by computerized programmed operations using a genetic algorithm.

15. (Amended) The method according to claims 9 or 10

wherein the training is automated by computerized programmed operations using a genetic algorithm reduced in computational complexity by including the steps of:

grouping alleles and/or characteristic SNP patterns into

families as are defined by (i) having similar expression patterns, or (ii) being turned on and off by another gene, or (iii) both having similar expression patterns and being turned on and off by the same gene; and

starting training of the neural network with the genetic algorithm by using the families so created as single inputs to the neural network, the training with the genetic algorithm continuing repetitively until, families of greater and lesser significance being identified, it becomes computationally possible to train the neural network to genomic data consisting of individual alleles and/or characteristic SNP patterns;

wherein partitioning of all alleles and/or characteristic SNP patterns into families permits training of the neural network in a hierarchy of stages, first to the families and only then to the individual alleles and/or characteristic SNP patterns.

functionally similar for a majority of clinical outputs yields a method of predicting the effects of given drugs on clinical outputs of interest, as described in the section entitled "Use for Prediction of Drug Efficacies." We use this method to predict the effect of each of a pair of drugs on a given clinical output. This clinical measure may be a drug efficacy measure: for example, a combination of the extent of reduction of problematic symptoms or of the lack of specified side effects. We then compare this clinical measure for a given patient for each of the two drugs. If the clinical measure is a cost of treatment (such as a financial cost or a measure of patient suffering from side effects), a drug minimizing this cost may be chosen.

6.5 Use for Choosing Optimal Drugs for a Given Patient

The above comparison of drug efficacies allows the development of an automated technique for choosing optimal drugs for a given patient. A given patient's genome is first scanned and the problematic genomic inputs (such as problematic alleles) identified

(as those elements of the genomic inputs that are also present in the universal functional categories). A software program then identifies which drug is expected to perform the best on the patient's set of problematic inputs. The program does this by comparing the effectiveness of different drugs on the problematic inputs found in the given patient.

7. Conclusion

In accordance with the preceding explanation it should now be understood that the present invention embodies new, neural-network-based, methods of identifying and relating particular alleles -- out of a vast number of alleles present in the genomic sequences of each of a large number of individual organisms -- that are relevant in a practical sense to (i) some particular biological or sociological problem, normally disease, afflicting or besetting the organisms, and, separately, to (ii) various therapies, normally drugs but also including environmental changes, that may be applied to the

/

EXHIBIT 14

Hypertension Response Prediction
SBIR Phase I
Final Report

Jan. 18, 2002

Sponsored by
National Science Foundation
SBIR/STTR Program

Prediction Sciences
Personalizing Medicine Today™

Table of Contents

INTRODUCTION

Project Summary

Our project is to use patient-specific characteristics and genetic variables, processed by neural nets, to predict response or non-response of the patient to anti-hypertensive therapy. The goal of our Phase I research was to demonstrate proof of principle for generating predictive power of patient-specific variables in a population of hypertensive patients, for response or non-response to firstline therapy.

Our efforts in this program were focused on 1) the development of a medication pathway-based multi-gene methodology for prediction, 2) the development of neural network processing towards the response/non-response outcome, and 3) the evaluation of predictive power in our system. Our long-term goal is to develop a system that clinicians can use, at the point of care, to prescribe anti-hypertensive therapy with the greatest likelihood of success for each individual.

Firstline anti-hypertensive medication was traditionally either a diuretic or beta-blocker. In the past decade, both Calcium Channel Blockers and ACE Inhibitors have been used increasingly, while the traditional medications have decreased. After much consultation with physicians, evaluating our data source, and further researching pathways, we decided to pursue ACE Inhibitor response or non-response for our Phase I proof-of-principle study.

Project Outcome

The generation of our ACE Inhibitor Responder/Non-responder pharmacogenomic model was completed and is detailed below. Objectives that were met and key accomplishments/findings are summarized below.

1. **Patient DNA and Medical data acquisition**

Key accomplishments:

- Over 1200 hypertension patients reviewed for study
- 100 patients selected
- Largest group (60) found responsive/nonresponsive to ACE inhibitor monotherapy

2. **Development of Medication pathway multi-gene system**

Key accomplishments:

- Comprehensive literature search for gene variants which affect hypertension medication response
- Twenty-two SNPs selected over three biochemical pathways
- SNP selection able to be updated twice over six-month study period
- Additional thirty SNPs selected for Phase II

3. **Development of pharmacogenomic neural network processing software**

Key accomplishments/findings:

- Pre-processing sorting and annotating SNP software created
- Advanced neural network routines developed on MATLAB environment
- Neural network algorithms validated on NCI AML/ALL gene expression data
- Novel Bayesian thresholding technique invented
- Novel Functional Partitioning Method invented for future neural network design

4. **Patient Genotyping**

Key accomplishments/findings:

- Primer Design developed for 150 SNPs
- SNaPIT™ technology shown to be efficient and inexpensive method of genotyping
- 100 patients genotyped

5. **Analysis of Results**

Key findings:

- Most individual SNPs selected have weak (<10%) to medium (<25%) correlation with hypertension R/NR
- Majority of patients homozygous for SNPs selected
- Stronger correlation with ACE R/NR is shown with aggregate groups of SNPs
- Stronger correlation with ACE R/NR is shown with aggregate groups of lifestyle variables
- Weak-Medium correlation of ACE R/NR with lifestyle habits (6-23%)
- Medium correlation of ACE R/NR with body mass and ethnicity (40%)
- Neural network pharmacogenomic model gave 80% correct predictive values for ACE R/NR, weighted towards responders
- Neural network medical chart model gave 73% correct predictive values for ACE R/NR, weighted towards nonresponders.
- Combined genetic and medical chart model performed best, 83% correct predictive values for ACE R/NR, equal for R and NR
- Neural method superior to statistical tests, i.e. Nearest Neighbor Discriminant Analysis

6. **Development of phase II milestones/commercialization plan**

Key accomplishments/findings:

- Phase II milestones identified
- Phase III commercialization plan developed
- Partners for Phase III financing under negotiation
- Continuing funding obtained for additional research prior to Phase I funding

TECHNICAL BACKGROUND

Blood Pressure

A normal blood pressure reading for adults is considered to be around 120/80 mmHg. The first number represents the systolic pressure (the pressure when the heart is contracting) and the second number represents the diastolic pressure (the pressure when the heart is relaxed). A high blood pressure is 140/90 mmHg or above. High blood pressure can be divided into two major categories. When high blood pressure occurs without apparent cause, it is known as primary or essential hypertension; and when it occurs because of another disease, such as poor kidney function, it is known as secondary hypertension. Anyone can have temporary high blood pressure, resulting from excitement, nervousness, exertion, anger, fatigue, cold or smoking. In hypertension, high blood pressure is sustained over a period of time.

Prescribing Practice Trends

A review of three recent studies that have focused on the discontinuation of initial antihypertensive medications, revealed that after 6 months, less than 50% of patients are on initial treatment, regardless of class (diuretics, beta-blockers, calcium channel inhibitors, ACE inhibitors). Anti-hypertensive medication noncompliance is common and leads to substantial morbidity for patients and increased heath care costs. The Joint National Committee on Hypertension (JNC) recommends protocol on detection, evaluation, and treatment of high blood pressure. Current protocol and practice offers no systematic method to take in account individual characteristics, phenotypic or genetic.

Genetics of Hypertension Drug Response

A variety of mechanisms determine drug response. The pharmacokinetics include drug absorption, distribution, excretion, and metabolism. While the polymorphisms of genes involved in these processes have been in the focus of pharmacogenetic investigation, pharmacogenetics of modern anti-hypertensive drug therapy requires a focus on effector pathways and target genes. Current hypertensive drugs are not predominantly metabolized by known polymorphic enzymes and have flat dose-response relationships. Between classes with considerable differences in pharmacogenetic properties, the magnitude of blood pressure lowering is similar.

The target pathway of many anti-hypertensive medications, including ACE inhibitors, is the Renin Angiotensin System. This is illustrated in Figure 1.

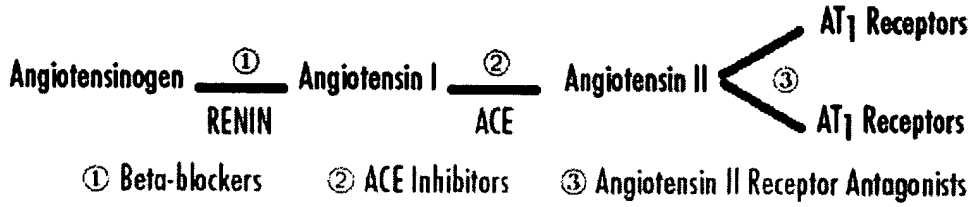

Figure 1: Renin Angiotensin System

Angiotensinogen, of the serine protease inhibitor family, is a cell-secreted plasma protein in the circulation originating predominantly from the liver. It is cleaved by Renin to release a small 10 amino acid protein, Angiotensin I. A positive feedback mechanism exists between Angiotensin II and Angiotensinogen expression.

Angiotensin Converting Enzyme (ACE), adipeptidyl carboxypeptidase, has 2 mechanisms for vasoconstriction. It converts angiotensin I to angiotensin II, which is a potent vasoconstrictor, and it inactivates bradykinin, which is a vasodilator. ACE contains one functional site for cleaving terminal dipeptides. There are different classes of ACE inhibitors. Non-peptide inhibitors chealate Zinc and heavy metal ions needed for enzymatic activity, and creates a catalytically defective enzyme. A second class of inhibitors are peptides that interact with ACE similarly to endogenous substrates. Most medications are in this class.

Angiotensin II receptors bind free angiotensin II and initiate the biochemical signal pathways that lead to many physiological effects downstream effects. It is a member of the superfamily of G protein-coupled receptors that have seven transmembrane regions. Exact mechanisms of signaling and regulation of expression are not known.

Neural Networks

Since the 1960's, neural networks have been applied to an expansive variety of problems. They are especially suited to practical applications that are complex and difficult to fully understand, but that have training data available. Their growing popularity in fields outside of computer science is rooted in their ability to train themselves to a given set of sample data. By simultaneously processing numerous independent variables in parallel, neural networks can assign different weights for each variable and can adjust these weights to varying situations. This adaptability allows more accurate, non-linear extrapolations of the data than do other forms of rigid mathematical models. In addition to parallel processing, another main advantage of neural networks over traditional, linear programs is that they can integrate the features of a non-comprehensive set of test cases and apply them to new but related cases.

Neural networks possess several qualities that render them amenable to the construction of predictors for genetic data. Foremost among these is that they are one of the only ways to model systems where a causal model is not available. This is certainly the case of

population genomics, where the significance of individual SNPs is typically unknown. Neural networks are also scalable to the incorporation of as yet unidentified factors, thus allowing the addition of new variables into preexisting models. Finally, once structured, the networks can adapt to nearly any problem; in this case to any medication or drug class.

Neural networks are essentially methods of performing a nonlinear least squares fit to a given data set. The data set consists of examples of correct input/output pairs, whose features the neural net will "learn" during training. The inputs may consist of continuous quantities, such as age, weight, height, or quantities derived thereof, or from *fuzzy inputs.* The latter may assume a continuous range of values, but only have well-defined meanings for a discrete set of values. For example, the property of being Hispanic could be modeled with a fuzzy input, with the discrete values 0 and 1 indicating the presence or absence of this property, but intermediate values could be used for a person with some Hispanic identity.

The form of the fitting function used for the least squares regression is generic, in the sense that a sufficiently large network can model a continuous function to any desired precision. The exact fitting function used is complicated by the fact that it incorporates many free parameters, multiple applications of a simple nonlinear function, and multiple linear combinations of intermediate quantities. The free parameters are set during training of the network; the nonlinear function repeatedly used (the *transfer function*) is typically a sigmoid function, which varies monotonically between 0 and 1; the multiple linear combinations combine the free parameters with the transfer function values.

The flexibility of neural nets to generically model data is derived through a technique of "learning." Given a list of examples of correct input/output pairs, a neural net is trained by systematically varying its free parameters (*weights*) to minimize its chi-squared error in modeling the training data set. Once these optimal weights have been determined, the trained net can be used as a model of the training data set. If inputs from the training data are fed to the neural net, the net output will be roughly the correct output contained in the training data. The nonlinear interpolatory ability manifests itself when one feeds the net sets of inputs for which no examples appeared in the training data. The net still produces output values in this case; they are based on features extracted from related input/output pairs in the training data. In this manner neural net "learns" enough features of the training data set to completely reproduce it, up to a variance inherent to the training data.

When using any predictor (not just neural networks), the use of distinct subsets of the available data for training and testing is required to ensure generalization. Usually, the training set is further broken down to distinct training and validation sets: the parameters of the predictor are set with respect to the training data set. For a training algorithm such as a neural network, the training occurs in successive stages. To avoid over-training (i.e., where the predictor memorizes specific features present in the training data set but not the trend; see Figure 2), this succession of training steps is discontinued when the error on the validation set begins to increase significantly. We use the error on the testing data set

as an estimate of how well we can expect our predictor to perform on new testing data as it becomes available.

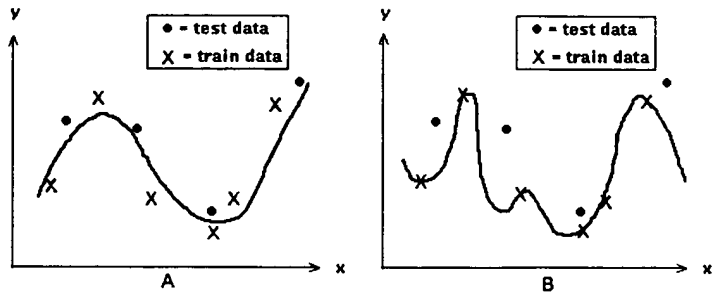

Figure 2. (A) A properly trained neural net will identify the trend inherent in the training set and fit not previously "seen" test data with minimal error. (B) An over-trained neural net memorizes the features of the training set and fits it with a very low error, but fits test data poorly.

Once optimal weights have been determined, the trained net can be used as a model of the training data set. If inputs from the training data are fed to the neural net, the net output will be roughly the output contained in the training data. The nonlinear interpolatory ability of the neural net manifests itself when one feeds inputs for which no examples appeared in the training data. Based on features extracted from related input/output pairs in the training data, the neural net will predict a response phenotype. The neural net "learns" enough features of the training data set to reproduce its trends.

EXPERIMENTAL PROCEDURE

Patient data extraction

We concentrated our SNP investigation on Angiotensin, ACE, and Type1 Angiotensin II receptor. We chose 21 SNP's total for our initial study. These three genes are the most closely involved in the effector pathways of ACE Inhibitor medication. Some recent studies have implicated SNP's in these genes in the response to ACE inhibitor medication. Pyrosequencing Inc. was granted a patent on some SNP's in this system. Table 1 shows genes considered for usage in this multigene study.

Table 1: Genes Involved in Hypertension to be Genotyped for SNPs.

Angiotensinogen lipoprotein lipase	α-adducin C3C complement component
Angiotensin II type I receptor Pancreatic phospholipase A2	β2-andrenergic receptor Haptoglobin
Angiotensin converting enzyme Insulin receptor, LDLR	Renin MNS Blood Group
α2-andrenergic receptor 6-phosphogluconate dehydrogenase	β1-andrenergic receptor Endothelin-1
Na+-H+ antiporter Human Leukocyte Antigen System	a2c10-andrenergic receptor.

Justification of the candidate gene strategy.
We have chosen a **candidate gene approach** as our primary strategy for this study, as opposed to a genome-wide approach, for several reasons:

A. Already documented success of candidate gene allelic association studies in complex human traits. As demonstrated by the now-classic example of association of one form of late-onset familial Alzheimer disease to a particular allele at the APOE locus (Martin et al, 2000), allelic association to particular alleles at candidate loci has emerged as a viable strategy for disease gene discovery. In the realm of drug responses, candidate gene allelic association studies have also been successful (Roses, 2000).

B. Potential for greater power for allelic association than non-parametric linkage in complex human traits. Risch et al (Risch, 2000. Risch & Teng, 1998, 1999) have simulated statistical power to detect complex trait predisposition genes. Depending upon a number of assumptions (heritability, marker spacing, marker heterozygosity) the allelic association approach may actually be more powerful (require fewer genotype assays in fewer subjects) that the non-parametric linkage approach in complex traits.

C. Current lack of feasibility of genome-wide allelic association studies. Allelic association studies are not currently feasible on a genome-wide basis. Genetic linkage, in families (parametric) or sib-pairs (non-parametric), is amenable to genome-wide approaches, since co-segregation of genotype and phenotype crosses one generation at a time, enabling genome-wide microsatellite marker spacing at wide intervals (customarily, 10 cM intervals, where 10 cM is likely to be ~10 Mb). By contrast, allelic association studies in unrelated individuals rely on linkage disequilibrium (LD; between marker and trait alleles) which may reach back for hundreds of human generations, depending upon when the ancestral LD between marker and trait alleles was established. At the human LPL (lipoprotein lipase) locus, re-sequencing (Nickerson et al, 1998) demonstrated that LD occurs over only very short ranges. Kruglyak (Kruglyak, 1999) estimated that population LD may extend for as little as 3-10 kbp genome-wide; if a 10 kbp is the standard, then genome-wide allelic association studies would require (~3 x 10^9

bp)/(1C x 10^3 bp)= ~3 x 10^5 SNP markers, a number not currently feasible. For example, commercially available human SNP chips (Affymetrix) are designed primarily for linkage rather than association, with only ~2000 SNP loci currently available, spacing far too sparse for LD.

Patient Chart Data

Patient Characteristic Variables
Age
Ethnicity
Height
Weight
Diet Score
Exercise Score
Smoker/NS
Alcohol Intake
Other Afflictions
Blood Pressure
Outcome Variables
Resulting BP
Side effects (if any)
Secondary treatment (if any)
Lifestyle Changes

Table 2 : Lifestyle variables to be considered in study

We obtain from patient medical charts the parameters listed in table 2. These parameters will complement the genotypic binary allele variables of the previous section. The complete set of parameters for each patient will then include those in table 1, the patient's genetic data, and the anti-hypertension treatment prescribed. This complete set of parameters will be used as inputs to the neural network. The output will consist of the observation of whether the therapy was considered successful: that is, whether a secondary therapy was subsequently prescribed.

CLASSIFICATION OF PATIENT RESPONSE/NON-RESPONSE

Normal blood pressure for an adult is around 120/80 mmHg. We defined a hypertensive patient as a patient with BP above 149/90 mmHg.

The simplest definition of a "Responder" is a patient who is hypertensive, takes an anti-hypertensive medication, and has consistent normotensive BP after medication treatment without adverse side effects.

We use the initial diagnosis of hypertension by the physician on the patient's chart as sufficient information to classify the patient as hypertensive. A BP at diagnosis, or a series of prior hypertensive BP measurements reinforces this diagnosis.

We set a minimum duration of 6 months since initial diagnosis of hypertension. We further reinforce this by a minimum 6 months duration of one medication therapy if a patient has switched medication.

To be classified as normotensive after medication, patients must have at least 3 normotensive BP measurements over a minimum 6-month duration of therapy, and no more than 1 hypertensive measurement after medication start date. Patients must also have no adverse side effects noted on charts during the duration of medication. If a patient was not quite normotensive, yet, doctor's discretion kept the patient on the medication for over a year, we also allowed responder classification as a 40+ mmHg reduction in systolic or diastolic blood pressure.

The simplest definition of a "Non-responder" is a patient who is hypertensive, takes an anti-hypertensive medication, and has continues to have a hypertensive BP after medication treatment, experiences adverse side effects, or must revert to alternative medication to attain normotensive BP measurements.

We do not distinguish between ACE Inhibitors in this early stage. We also do not distinguish non-response due to lack of BP reduction vs. non-response due to adverse effects.

GENOTYPING

We utilized Hibergen Inc.'s SNaPIT^{TM} Technology for the detection of our chosen SNPs. This technology allowed us several accuracy, time, and cost advantages. Accuracy is optimal because SNP detection is not dependant on heteroduplexing of DNA. Instead SNP's are detected using Mass Spectrometry sequencing techniques. Insertion-deletion mutations and homozygous/heterozygous detection are all detectable with a single designed assay.

For each SNP chosen, sequence information was obtained using GenBank search of the gene for which the SNP was chosen. Sequence is cross-verified by 2 parties. PCR (polymerase chain reaction) primers flanking the SNP are designed to amplify a 50-80 base pair fragment. Designed primers pairs are optimized for annealing temperature, lack primer dimerization, and lack of primer self-complementation. PCR amplification is tested for robustness and specificity.

Experimental amplification is done with modified nucleotides. These modified nucleotides are specifically recognized by highly specific DNA Glycosidase enzymes which cleave the product at those positions. The resulting fragments are analyzed by sequencing mass spectrometry methods. The sequence reveals the genotype of the DNA for the chosen SNP.

RESULTS AND ANALYSIS

VALIDATION OF NEURAL NET

A neural network is trained on data obtained from independently diagnosed patients and the resultant model is used to predict a new patient's response. In an attempt to demonstrate that genetic data can successfully be used to train a neural network to predict a phenotypic outcome, a network was trained on a publicly available data set. AML/ALL data of Golub et al. Contains the expression levels of approximately 6000 genes. Golub identifies 50 genes that correlate most highly with the ALL/AML class distinction. For this report, only these 50 genes were used.

In our hypertension experiment, we use SNP data as inputs for the neural net. For each patient, the data would be a sequence of 1's and 0's depending on whether or not a person has a particular mutation. Since the AML/ALL dataset is not binary, a method of transforming the continuous expression data to binary data is needed. To perform this transformation, a bayesian approach is taken. The reasoning behind this approach is as follows: The expression of a particular gene in a patient indicates the presence or absence of that gene. If the expression is low, the gene is not present, if the level is high, the gene is present. Hence, the distribution of expression levels from a gene that correlates with a particular phenotype classification is bimodal. Strictly this is the case only for a phenotype that is bimodal and it has not been proved. (ex eye color). Further, the on/off distinction is false and is merely a high/low difference (i.e. if the gene is expressed, it is already on). In addition, a close examination of the data suggests that the distribution may be tri-modal for some genes. (there is not enough patient data to verify this). In general the distribution appears as a Gaussian peak anywhere between 0-1000 and a very long tail out as far as 15000 with lumps in the tail. The idea is the data is distributed as a mixture of two Gaussians, one "sharply" peaked near the origin and a second more diffuse peak in the tail. In the Bayesian approach we model the data is follows. The probability for a patient's gene expression for the same gene, say G1, is modeled as

$$p(x|\theta) = \Sigma_{i=1,2}\, p(i)\, p(x|i,\theta)$$

In fig. 1, an example of the results of this Bayesian analysis is presented for the gene MCL1. The bar charge is a histogram of the expression levels above 1497 were classified as On, below is Off. This classification correlated well with the AML/ALL distinction, with a correlation coefficient of 64. The ON correlated with the ALL, off with AML. There were 60/72 matches.

The bayesian routine above transforms the expression data to "genetic data". The genetic data is now passed to a neural net. No additional pre-processing is done. The data is split into training, test, and validation subsets. The network has 51 (number of genes +1) hidden neurons and one output neuron. The hidden layer used a "tansig"activation function and the output layer uses a "log sig function. After training, the net is simulated

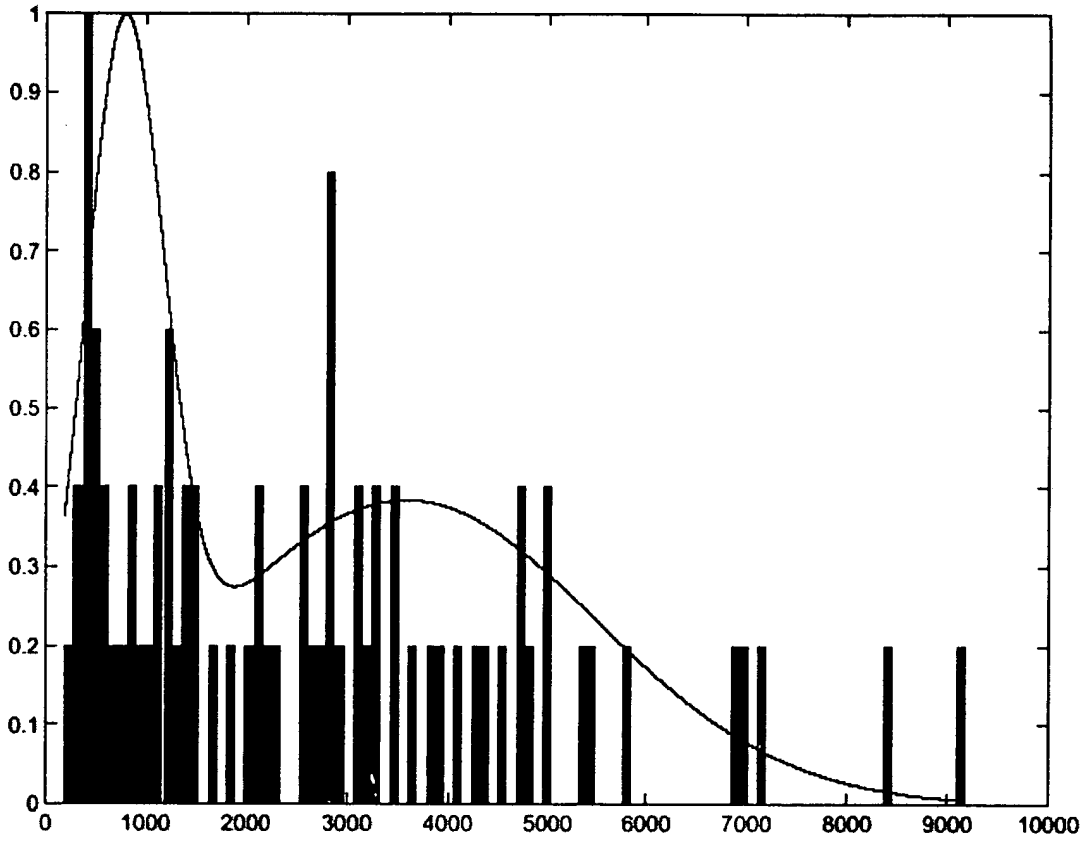

Figure 3: The resulting probability distribution for gene MCL1. Expression levels above 1497 were classified as on below as off.

again using the test data alone. The output of the network is fit to the target using a standard, linear regression analysis.

Regression analysis was plotted for the trained network. The linear fit gives a correlation coefficient R= 0.999 indicating good agreement. With a threshold of +- 0.3, we conclude that the neural network made correct predictions for 100% of the (test) patients. This indicates that the network generalizes well to new data.

ASSOCIATION STUDY FOR ACE INHIBITORS

We performed an association study to determine the ability to predict response/non-response of hypertension patients to ACE Inhibitors. For this study, we wished to determine whether or not a patient's genotype could be used a predictor of response/non-response. To this end, we selected 20 candidate SNP locations in the ACE system for genotyping. The patients were genotyped by Hibergen of Dublin, Ireland. Both copies of the patient's genes were discovered. Due to a limited sample population, allelic frequencies were not identified. This study consists of a group of efforts to determine whether or not we could predict a patient's response/non-response given his/her genotype. We first give frequencies of occurrence of a genotype at the given SNP location and then calculate the degree of correlations between genotype and phenotype. We use the top correlates to perform a cross-validated, Nearest-Neighbor discriminant analysis prediction.

Frequency/Correlation Study

We first discovered the frequency of genotype in the sample.
An explanation is in order. If the genotype at the genetic position is G and a patient can have G/G, G/A or A/G, A/A at that position, then within the responder or non-responder sample population, we measure the fraction of homozygous G /G, the fraction of heterozygous A/G, and the fraction of homozygous A/A. Suppose we have 5 patients with genotypes (L/R)=G/A, GG, A/G, A/A, GG at location 1 and phenotypes R, R, NR, NR, NR. Then for location 1, we report 6 numbers: the fractions of responders that are homozygous G, heterozygous, and homozygous A, and the fractions of non-responders that are homozygous G, heterozygous and homozygous A. The results for
The 20 SNP locations are presented in table 3. Note that for responders, f ho-L + f he + f ho-R $\equiv$ 1. A similar equation holds for the non-responders.

Because the sample population is limited to fewer than 60 patients, Table 3 should not be interpreted as provided a comprehensive study of allelic frequencies in the population at large. It does give some indication and guide to which SNP location might serve as a good classifier. In order for the SNP location to be a good classifier, the distribution of responders and non-responders over the 3-genotype classes should not be too similar. For example, PS5 would not be a good classifier since all responders and non-responders are homozygous GG. On the other hand, PS4 might serve as a good classifier, since %94 of the non-responders are homozygous GG, but only %61 percent of the responders share that genotype.

Table 3: Frequency f allele Resp nse/Nonresponse in patient population

Allele label	Fractions			
	Responder/ Non-Responder	Ho-L	He	Ho-R
PS1(T/A)	R	0.94	0.06	0.00
	NR	0.88	0.12	0.00
PS2(G/A)	R	0.52	0.39	0.10
	NR	0.94	0.06	0.00
PS4(C/T)	R	0.61	0.32	0.06
	NR	0.94	0.06	0.00
PS5(G/A)	R	1.00	0.00	0.00
	NR	1.00	0.00	0.00
PS6(G/A)	R	0.52	0.39	0.10
	NR	0.88	0.12	0.00
PS7(G/A)	R	0.97	0.03	0.00
	NR	0.94	0.06	0.00
PS8(C/A)	R	0.00	0.16	0.84
	NR	0.00	0.29	0.71
PS10(C/T)	R	0.90	0.10	0.00
	NR	0.76	0.18	0.00
PS9(G/A)	R	0.06	0.52	0.39
	NR	0.47	0.24	0.35
PS13(G/A)	R	0.00	0.13	0.87
	NR	0.06	0.18	0.76
PS18(G/A)	R	0.97	0.03	0.00
	NR	0.94	0.06	0.00
PS19(C/T)	R	0.23	0.55	0.23
	NR	0.29	0.41	0.29
PS20(G/A)	R	0.16	0.68	0.16
	NR	0.18	0.53	0.29
PS22(G/A)	R	0.74	0.00	0.26
	NR	0.71	0.06	0.24
PS23(C/T)	R	0.16	0.68	0.19
	NR	0.18	0.53	0.29
PS24(G/A)	R	0.90	0.06	0.00
	NR	0.94	0.00	0.00
PS31(G/A)	R	0.03	0.45	0.52
	NR	0.00	0.29	0.71
PS37(G/T)	R	0.84	0.10	0.06
	NR	0.71	0.29	0.00
PS40(G/A)	R	0.68	0.29	0.03
	NR	0.65	0.24	0.12
PS49(G/A)	R	0.90	0.10	0.00
	NR	0.94	0.06	0.00
PS50(C/T)	R	1.00	0.00	0.00
	NR	1.00	0.00	0.00

In principle, one could classify a patient's phenotype with a single SNP location if there was a high correlation between phenotype and genotype at that location. Then, one could use a simple association rule to discriminate between responder/non-responder populations. Hence, we calculated Pearson's r, the linear correlation coefficient, for each genotype. This information is in Table 4. Pearson's r gives a number between -1 and 1, with -1 indicating a high negative correlation, 1 a high positive correlation and 0 no correlation. In case a variable is constant(no genotypic variation), Pearson's r is undefined. We are only interested in the magnitude of the correlation, so the absolute value $|r|$ is presented in table 4. We also supply the significance of the correlation, P. P is the complimentary error function and it gives the probability that $|r|$ should be larger than its observed value in the null hypothesis that the variable is uncorrelated with response/non-response. In other words, P is the probability that if the variable is uncorrelated then $|r|$ would be larger than what is observed $|r|$. A small value of P indicated that the variable is significantly correlated with response/non-response.

As a guide, one would like to interpret the linear correlation function as follows: if the correlation is 1, then we could predict with certainty the phenotype of the patient, but if the correlation is 0.5, then we could predict half of the patients correctly. As table 4 indicates, no genotypes at any SNP locations correlate with phenotype better than 0.5. Since no single genotype correlates well, we want to consider combinations of genotypes as a single variable. We turn to this in the next section.

Note: While Pearson's r is typical used for continuous variables; it is not unreasonable to use Pearson's r in the case. The linear correlation coefficient is the dot product of the genotype and phenotype, when viewed as vectors in some high dimensional space. Geometrically, if the vectors are nearly aligned or anti-aligned, then the coefficient $\cong 1$ and the correlation is high. If the vectors are nearly perpendicular, then the coefficient $\cong 0$ and the correlation is low. Thus this geometrical interpretation justifies using Pearson's r as an imprecise, yet intuitive, indication of the correlation between genotype and phenotype.

Chi-Squared Statistic

We also performed a chi-square test on the data. Chi-square gives a measurement of association. The null hypothesis is that two variables have no association. A high chi-square value indicates that the null hypothesis is unlikely. The chi-square statistic is given by

$$\chi^2 = \sum_{i,j} \frac{(N_{i,j} - n_{i,j})^2}{n_{i,j}}.$$

In our case, we are measuring the association between phenotype and genotype. There are two phenotype values, response/non-response, and three-genotype variable, 2 homozygous combinations and one heterozygous combination. Here, $N_{i,j}$ is the observed number of responses/non-responses for each genotype value. n_j is the number of responses/non-response in the null hypothesis.

In table 5, we present the chi-square value and it's significance. The significance is given by $Q(\chi|\nu)$, an incomplete gamma function, where ν is the degrees of freedom. Strictly speaking, Q is the probability that the sum of the squares of ν random normal variables of unit variance (and zero mean) will be greater than χ^2. Essentially, we are asking whether or not the sum of errors, N-n, are less than the sum of errors from an uncorrelated distribution of random variables. If Q is high, then it is likely that the errors of the uncorrelated distribution are smaller greater than the errors in the data, indicated that the data is likely more correlated than some multi-variate normally distributed data set.

Nearest-Neighbor Study

In this section we use an aggregate data vector as a discriminator. Since no genotype correlates above 0.5, we attempt a multivariate, nearest neighbor discriminant analysis of the data. The nearest neighbor discriminant analysis is a nonparametric procedure. While it does not depend on the distribution of the data, it does depend on the notion of the distance between pairs of observations. The basic idea of the NNDA (nearest neighbor discriminant analysis) is as follows: To classify a new observation, find the observation in the calibration set that is nearest and assign the new observation to be in the same class as that which the nearest observation comes from.

Since our sample population is small, we use a leave-one-out cross-validation procedure to test all samples. To carry out this procedure, each sample is removed from the population, one are a time. The remaining samples are used as the calibration set. The left out sample is then tested using the NNDA and assigned to a class. The class distinction is response/non-response. The aggregate data vector is the collection of SNP location genotypes that correlate best with the class distinction. In case of a tie (see Table 4), the genotype was selected randomly.

In Table 6 we present the results of the leave-one-out, NNAD for a variety of aggregate vectors. Column one gives the list of SNP location genotypes and the remaining columns give the percent of correct predictions and the percent of incorrect predictions. Each sample is predicted. At this time, we do not have a metric to indicate the strength of an individual prediction. To match the aggregate vector in table three, read the row vector of table two that corresponds to the numbers in the aggregate vector. For example in table 6, the aggregate vector that gives the greatest prediction strength (%76 correct) is V = {25,4,7,13). This corresponds to the aggregate genotype {PS4 ho-L, PS6 ho-L, PS10 he, PS11 ho-L}. The biological implications should be discussed, perhaps.

Note that in table 6, the topmost column includes a genotype from each SNP location. As one proceeds down the table, lower correlates are pared from the aggregate variable until, in the last row, only the most highly correlating genotype from table 2 remains. Table 6 suggests that several genotypes add irrelevant information concerning class distinction. Adding SNP that correlate below the .2 level tend to decrease the prediction strength of the aggregate vector. While PS9 alone discriminates at a level equivalent to the aggregate vector V = (25,4,7,13), this may be an artifact of the small sample size. We

recommend that it would be better to proceed with caution and use the subset of SNP's that correlate above 0.25 to build a classifier. It may be a good idea to cut out the table below the vector (25,4,7,13). Clearly it is not necessary to use all SNP's but certainly we don't want to predict based on a single SNP.

Neural Network

We next presented the data to a neural network to see if the network could learn any association in the data between the response/non-response phenotype and a genotype containing a selected list of SNPs. We used a net with a single hidden layer. The output layer had a logarithmic sigmoid function. After training the net, a new patient sample was presented to the net for predicted. With the log-sig transfer function in the output layer, the net gave a value between 0-1. This value was interpreted as a probability, with a value near one indicating a high probability of patient response and a value near zero indicating a low chance of response. To reduce the number of false negatives, only network outputs that give a clear prediction are counted. This means result between 0.7-1 counts as a prediction that the patient is a responder and a result between 0-0.3 counts as a prediction for non-response. Due to the small sample population, we used a leave-one-out cross-validation procedure, similar to the procedure used for the Nearest-Neighbor discriminant analysis. Each sample was removed from the population, in turn. After the net was trained to the remaining data, the removed sample was then presented to the net for prediction.

The results are presented in table 7. The net performed best with the aggregate vector, (22,4,7,13,20). With this vector the net made clear predictions for %77 of the sample population with a %67 percent success rate. We expect to see significant improvement when we get our hands on more patients and take some blood from them.

All results in table 7 are for a net with twice as many hidden nodes as the number of SNP's presented as classifiers. Using this aggregated vector, the network was optimized to achieve the highest sensitivity level, i.e. to make the most prediction. This optimization was done by changing the number of nodes in the hidden layer. The highest sensitivity achieved was %80. The accuracy, however, was only marginally improved and did not go above the %68 percent level.

Table 4: Allele c rrelations with ACE inhibitor response/nonresponse

Table 4	Correlations			
Label	SNP Location	Genotype	Correlation	Significance
1	PS1	T/T	0.09	0.53
2		T/A	0.09	0.53
3		A/A	~	~
4	PS2	G/G	0.43	0.00
5		G/A	0.37	0.01
6		A/A	0.16	0.28
7	PS4	C/C	0.34	0.02
8		C/T	0.31	0.03
9		T/T	0.11	0.45
10	PS5	G/G	~	~
11		G/A	~	~
12		A/A	~	~
13	PS6	G/G	0.36	0.02
14		G/A	0.30	0.05
15		A/A	0.16	0.28
16	PS7	G/G	0.07	0.66
17		G/A	0.07	0.66
18		A/A	~	~
19	PS8	C/C	~	~
20		C/A	0.16	0.28
21		A/A	0.16	0.28
22	PS10	C/C	0.12	0.43
23		C/T	0.12	0.43
24		T/T	~	~
25	PS9	G/G	0.50	0.00
26		G/A	0.29	0.05
27		A/A	0.03	0.83
28	PS13	G/G	~	~
29		G/A	0.07	0.66
30		A/A	0.07	0.66
31	PS18	G/G	0.07	0.66
32		G/A	0.07	0.66
33		A/A	~	~

Table 4(Continued)

Table4	Correlations			
Label	SNP	Genotype	Correlation	Significance
34	PS19	C/C	0.08	0.61
35		C/T	0.08	0.61
36		T/T	0.01	0.95
37	PS20	G/G	0.02	0.90
38		G/A	0.13	0.39
39		A/A	0.16	0.28
40	PS22	G/G	0.01	0.95
41		G/A	0.20	0.17
42		A/A	0.06	0.68
43	PS23	C/C	0.02	0.90
44		C/T	0.13	0.39
45		T/T	0.12	0.43
46	PS24	G/G	0.07	0.64
47		G/A	0.16	0.28
48		A/A	~	0.00
49	PS31	G/G	0.11	0.45
50		G/A	0.17	0.27
51		A/A	0.20	0.19
52	PS37	G/G	0.16	0.28
53		G/T	0.26	0.08
54		T/T	0.16	0.28
55	PS40	G/G	0.03	0.84
56		G/T	0.06	0.67
57		T/T	0.17	0.24
58	PS49	G/G	0.07	0.64
59		G/A	0.07	0.64
60		A/A	~	0.00
61	PS50	C/C	~	0.00
62		C/T	~	0.00
63		T/T	~	0.00

Table 5: Chi-squared analysis of data

Table 5	Chi-Squared Analysis	
SNP Location	Chi-Squared Statistic	Q Significance
PS1	0.41	0.03
PS2	8.99	1.00
PS4	6.05	0.98
PS5	0.00	0.00
PS6	6.66	0.99
PS7	0.19	0.01
PS8	1.17	0.18
PS9	10.45	1.00
PS10	0.78	0.09
PS13	2.14	0.45
PS18	0.19	0.01
PS19	0.82	0.10
PS20	1.33	0.22
PS22	1.87	0.37
PS23	0.88	0.11
PS24	1.12	0.16
PS31	1.91	0.38
PS37	3.91	0.84
PS40	1.42	0.24
PS49	0.21	0.01
PS50	0.00	0.00

Table 6: Leave One Out-Cross Validated N-N Discriminant Analysis

Table 6 NNDA			
SNP Locations	Minimum Corr-Coeff	Percent Correct	Percent Incorrect
25,4,13,7,53,41,51,57, 39,21,47,44,22,2,34,58, 30,31,16,61,	0.0	0.62	0.38
25,4,13,7,53,41,51,57, 39,21,47,44,22,2	0.9	0.62	0.38
25,4,13,7,53,41,51,57	0.17	0.69	0.31
25,4,13,7,53	0.26	0.76	0.24
25,4,13,7	0.26	0.78	0.22
25,4,13	0.36	0.78	0.22
25,4	0.42	0.78	0.22
25	0.50	0.78	0.22

Table 7: Results of Neural Netw rk pharmacogenomic modeling

Table 7 Neural Net					
SNP Locations	Minimum Corr-Coeff	Percent Correct	Percent Incorrect	Percent R Correct	Percent NR Correct
25,4,13,7,53,41,51,57, 39,21,47,44,22,2,34,58, 30,31,16,61,10	0.0	0.60	0.40		
25,4,13,7,53,41,51,57, 39,21,47,44,22,2	0.9	0.62	0.38		
25,4,13,7,53,41,51,57	0.17	0.69	0.31		
25,4,13,7,53	0.26	0.76	0.24	0.89	0.50
25,4,13,7	0.35	0.80	0.20	0.97	0.50
25,4,13	0.36	0.80	0.20	0.93	0.56
25,4	0.42	0.78	0.22	0.93	0.50
25	0.50	0.78	0.22	0.93	0.50

Breakdown of correct values are in the final two columns to the neural net table. They give the percent of responders/non-responders correctly predicted. It looks like the aggregate vector (25,4,13) performs best, with %93 of the responders correctly predicted and %56 percent of the non-responders correctly predicted. Although you might argue that (25,4,13,7) did best with %97 and %50 respectively.

Also, while the Nearest Neighbor discriminant analysis appears to have done just as well as the net, it is not as robust a predictor as the net. This is because the NN discriminant analysis does not do any averaging, it simply finds the neighbor closest to the test case and assigns the test case to the nearest neighbor's own case. A more robust NN discriminant analysis might look for the 5 closest neighbors. In our case, since we have so few samples and more responders than non-responders, it seems that the non-robust NN discriminant analysis is more appropriate. Also, with regards to a the notion of distance in the NNDA. The distance between two samples is the following sum

$$D(i,j) = \sum_{SNP's} | SNPG_i - SNPG_j |$$

Here, j labels the patient and SNPGj labels the genotype. Since SNPGj = {0,1}, the distance function counts how many different SNPs two patient's have.

Medical Chart Data Variable Modeling

The above analysis was done on medical chart data as well. Several variables such as age, weight, and height were coarse-binned because of sample size and biological considerations. This is shown in table 8.

Table 9 shows that lifestyle variables had little correlation upon outcome, while there was a medium level of correlation of physical variables upon outcome. The majority of the population was Caucasian and gender seemed to make only a marginal difference upon response but not non-response.

Table 8: Medical Chart data frequencies

	Table 1	Frequencies		
	Diagnostic	Class	Responders	Non-Responders
1	Alcohol		0.548	0.824
2	Caffeine		0.871	1.000
3	Low sodium		0.258	0.235
4	Low fat		0.161	0.235
5	Exercise		0.613	0.882
6	Smoker		0.161	0.176
7	Gender	Male	0.419	0.529
		Female	0.581	0.471
8	Height	-60	0.129	0.000
9		61-63	0.194	0.059
10		64-67	0.323	0.647
11		68-70	0.258	0.118
12		71-73	0.097	0.118
13		74-77	0.000	0.000
14		78-80	0.000	0.059
15	Weight	-50	0.000	0.000
16		51-100	0.000	0.000
17		101-150	0.258	0.000
18		151-200	0.516	0.588
19		201-250	0.226	0.235
20		251-300	0.000	0.000
21		300-	0.065	0.059
22	Body	1.5-2.0	0.032	0.000
23	Mass	2.0-2.5	0.226	0.118
24		2.5-3.0	0.419	0.529
25		3.0-3.5	0.194	0.176
26		3.5-4	0.032	0.059
27		4-	0.065	0.000
28	Race	B	0.161	0.118
29		W	0.774	0.765
30		H	0.032	0.118
31	Age	-49	0.097	0.118
32		50-59	0.226	0.235
33		60-69	0.226	0.353
34		70-79	0.290	0.118
35		80-	0.161	0.176

Table 9: C rrelati n and Q-value f Medical Chart data

	Table 8			
	Diagnostic	Class	Correlation	Significance
1	Alcohol		0.033	0.871
2	Caffeine		0.035	0.865
3	Low sodium		0.146	0.476
4	Low fat		0.324	0.112
5	Exercise		0.174	0.394
6	Smoker		0.035	0.865
7	Gender	Male	0.092	0.653
		Female	0.092	0.653
8	Height(in.)	-60	0.193	0.343
9		61-63	0.287	0.160
10		64-67	0.130	0.525
11		68-70	0.008	0.967
12		71-73	0.138	0.498
13		74-77	~	~
14		78-80	0.325	0.111
15	Weight(lbs)	-50	~	~
16		51-100	~	~
17		101-150	0.329	0.107
18		151-200	0.015	0.939
19		201-250	0.041	0.841
20		251-300	~	~
21		300-	0.325	0.111
22	Body	1.5-2.0	~	~
23	Mass	2.0-2.5	0.008	0.967
24		2.5-3.0	0.222	0.276
25		3.0-3.5	0.041	0.841
26		3.5-4	0.325	0.111
27		4-	~	~
28	Race	B	0.041	0.841
29		W	0.103	0.612
30		H	~	~
31	Age	-49	0.325	0.111
32		50-59	0.243	0.235
33		60-69	0.130	0.525
34		70-79	0.370	0.070
35		80-	0.265	0.195

Table 10 and 11 show nearest neighbor and neural network predictive performance using medical chart information. Modeling shows that prediction involving several variables gives the best prediction, even though some of these variables are correlated with each other, such as body mass and height. One can conclude from this small population that non-genomic variables are as important as genomic variables in predicting response to ACE inhibitors. False positives and false negatives were not broken down because of the small sample size.

As with genomic variables as inputs, the neural network model was able to predict responders better than non-responders. This is probably because there were twice as many responders (39) than there were non-responders (19). However, in this case non-responders were predicted to a much greater rate. On a third of the patient population sodium and low fat information was judged to be incomplete. Thus these variables are thrown out of the analysis, even though they have high correlate values with R/NR. In table 11, a net was trained to only those who have this information for one of the variable sets, and the corresponding data is in parenthesis. The results are markedly better, thus for future studies we will pay particular attention to obtaining this information from patients.

Table 10: Medical Chart Data Nearest Neighbor Discriminate Analysis

Variable number	NNDA		
	Min. Corr	% Correct	%Incorrect
All	0.0	0.67	0.33
34,17,31,26,21,14, 4,9,35,32,24,8,5,3, 12,10,33,29,7	0.09	0.71	0.30
34,17,31,26,21,14, 4,9,35,32,24,8	0.1935	0.7083	0.2917
34,17,31,26,21,14, 4,9	0.2870	0.8333	0.1667
34,17,31	0.32	0.6250 (67%predicted)	0.3750
34,17	0.33	0.5882 (71%predicted)	0.4118
34	0.37	1.0(25%predicted)	0.0

Table 11: Neural Network Modeling of Medical Chart Data on ACE Response/Nonresponse

Variable number	Neural Net				
	Minimum Correlation	Percent Correct	Percent Incorrect	Responders Correct	Non-Responders Correct
All	0.0	0.5417	0.4583	0.6471	0.2857
34,17,31,26,21,14, 4,9,35,32,24,8,5,3, 12,10,33,29,7	0.09	0.6250	0.3750	0.8235	0.1429
34,17,31,26,21,14, 4,9,35,32,24,8	0.1935	0.6667	0.3333	0.7647	0.4286
34,17,31,26,21,14, 4,9	0.2870	0.7333 (0.8333)	0.2667 (0.1667)	0.7512 (0.9412)	0.7014 (0.5714)
34,17,31	0.32	0.6250	0.3750	0.8824	0.0
34,17	0.33	0.4167	0.5833	0.5882	0.0
34	0.37	0.71	0.29	1.0	0.0

Combined Medical Chart and SNP Data Variable Modeling

In this section we combine the genetic and physiological data in our pharmacological patient model of response. This report does not include sodium and low fat in the analysis, in which if we had complete information our models would appear to perform better (see previous section.) It appears that the aggregate vector (25,4,13,7,71,8) performs best. This combination is, from table 12, PS9,PS2,PS6,PS4, and height (which relates to body mass). The combination of genetic and medical information delivers better modeling performance overall, regardless of number of variables combined.

Overall, it appears that the genetic data somehow is able to find the responders and the medical data is able to find the non-responders. Further investigation in a larger sample set is required, which is why we have applied for Phase II funding. If these findings hold up in such a study, then a product generated from that research will have immediate and widespread application in the general population.

Table 12: Combined Genomic and Medical Variables

				Table 1	Labels				
1	PS1	T/T		34	PS19	C/C	64	Alcohol	
2		T/A		35		C/T	65	Caffeine	
3		A/A		36		T/T	66	Exercise	
4	PS2	G/G		37	PS20	G/G	67	Smoker	
5		G/A		38		G/A	68	Gender	Male
6		A/A		39		A/A			Female
7	PS4	C/C		40	PS22	G/G	69	Height(in.)	-60
8		C/T		41		G/A	70		61-63
9		T/T		42		A/A	71		64-67
10	PS5	G/G		43	PS23	C/C	72		68-70
11		G/A		44		C/T	73		71-73
12		A/A		45		T/T	74		74-77
13	PS6	G/G		46	PS24	G/G	75		78-80
14		G/A		47		G/A	76	Weight(lbs)	-50
15		A/A		48		A/A	77		51-100
16	PS7	G/G		49	PS31	G/G	78		101-150
17		G/A		50		G/A	79		151-200
18		A/A		51		A/A	80		201-250
19	PS8	C/C		52	PS37	G/G	81		251-300
20		C/A		53		G/T	82		300-
21		A/A		54		T/T	83	Body	1.5-2.0
22	PS10	C/C		55	PS40	G/G	84	Mass	2.0-2.5
23		C/T		56		G/T	85		2.5-3.0
24		T/T		57		T/T	86		3.0-3.5
25	PS9	G/G		58	PS49	G/G	87		3.5-4
26		G/A		59		G/A	88		4-
27		A/A		60		A/A	89	Race	B
28	PS13	G/G		61	PS50	C/C	90		W
29		G/A		62		C/T	91		H
30		A/A		63		T/T	92	Age	-49
31	PS18	G/G					93		50-59

32		G/A						94		60-69
33		A/A						95		70-79
								96		80-

Table 13: Nearest-Neighbor Discriminant Analysis on all Variables

Variable number	NNDA			
	Min. Corr	%Predicted	% Correct	%Incorrect
All	0.0	1.0000	0.5714	0.4286
Top 48 corre lates	0.875	1.0000	0.6190	0.3810
25,4,5,13,7,71,8,78,26, 14,64,53,69,72,91,75, 41,39,51	0.1840	1.0000	0.6429	0.3571
25,4,5,13,7,71,8,78,26, 14,64,53	0.2438	1.0000	0.6429	0.3571
25,4,5,13,7,71,8,78,26, 14,64	0.2831	1.0000	0.7143	0.2857
25,4,5,13,7,71,8	0.3234	1.0000	0.8333	0.1667
25,4,5,13,7	0.3558	1.0000	0.7619	0.2381

Table 14: Neural Network Performance on all variables

Variable number	Neural Net				
	Minimum Corre lation	Percent Correct	Percent Incorrect	Responders Correct	Non-Responders Correct
All	0.0	0.6190	0.3810	0.7308	0.4375
Top 48 correlates	0.875	0.6905	0.3095	0.8077	0.5000
25,4,5,13,7,71,8,78,26, 14,64,53,69,72,91,75, 41,39,51	0.1840	0.6667	0.3333	0.7308	0.5625
25,4,5,13,7,71,8,78,26, 14,64,53	0.2438	0.6905	0.3095	0.8077	0.5000
25,4,5,13,7,71,8,78,26, 14,64	0.2831	0.6905	0.3095	0.6923	0.6875
25,4,5,13,7,71,8	0.3234	0.8333	0.1667	0.8462	0.8125
25,4,5,13,7	0.3558	0.7619	0.2381	0.9231	0.5000

CHALLENGES

Our biggest hurdle was obtaining the patient data source. Cost restraints kept us from ideally generating a patient pool designed for our study. We relied on the archived data of a previous contract research organization's hypertension study, purchased by Pharsight Corporation (see evaluation of collaborators below). We purchased only the samples that passed our stringent criteria for classifying whether or not a patient responds to their medication. Examination of the Pharsight data quality and literature search on SNP's dictated that ACE Inhibitor therapy be our first choice for an initial prototype. Most medical data, even that in electronic form, is highly unorganized and requires significant pre-processing to be useful. However, we have lined up sources with high-quality data for Phase II and will be pre-purchasing such prior to Phase II funding among other reasons to develop data-specific pre-processing routines and thus speed up the path to commercialization. (see Development of phase II milestones/commercialization plan)

EVALUATION OF COLLABORATIONS

Data Source: Pharsight Corporation

Pharsight Corporation was newly developing their (data source supplying) branch of their company during our Phase I research. The medical records had not been put into electronic form as had been claimed and thus we were forced to sort through over 1200 patient charts by hand, deciphering hand-written notes, cross-referencing outcomes, etc. to obtain the patients that met our protocol for inclusion in this study. In addition, since we were the first group to purchase samples from their genomic DNA patient database, barriers during contract negotiation and finalization, sample extraction, and shipping, delayed our obtaining of the samples as well. These factors delayed our genotyping task by approximately 3 months, however, we successfully surmounted this obstacle to bring this study to conclusion on-time.

Our experience in overcoming these obstacles will greatly increase our efficiency in future research on this project and projects similar. We streamlined our protocol for extracting samples that meet our stringent criteria within Excel and Microsoft Access. We can better pre-evaluate archived data sources for future studies, based on the variable categories investigated in the study. Furthermore, we have gained much knowledge about hypertension research and clinical prescribing practices, so that when we conduct our own trial study to obtain data we are highly effective. We have already negotiated to obtain data for future research from three new institutions.

Genotyping: Hibergen Ltd.

The delay in obtaining samples severely limited the amount of time for genotyping and analysis. Our teams separately found the sequences of the SNP surrounding sequences and cross validated primer choice for accuracy. The genotyping technology gave us several advantages, and we will continue our relationship with Hibergen.

INTELLECTUAL PROPERTY

We will be patenting the functioning partition neural network methodology developed under this proposal, as well as the Bayesian method of gene stratification. We also plan to publish our results pertaining to ACE hypertension response/nonresponse pending verification in a larger sample set. Finally, we plan to patent specific genotypes following a further Phase II validation study.

IMPACT

On Prediction Sciences

By performing this study, we have provided further evidence that pharmacogenomic response in hypertension medications is multigenic and is well adapted to modeling by neural networks. We have used this study to attract outside investors, who, pending validation in Phase II of our results, have a high chance of investing in the company. In addition, we are negotiating with companies such as Quest Diagnostics and Pyrosequencing for partnering for commercialization of an eventual diagnostic.

On state

Prediction Sciences provided validation for SNAPIT technology, which has been exclusively licensed by Sequenom, a local San Diego company, and thus will help them sell more of their genotyping systems.

On nation

Prediction Sciences believes that development of a hypertension pharmacogenomic diagnostic will contribute significantly to curbing the rising cost of medical treatment in this country. Hypertension is prevalent among the elderly, which are the fastest-growing segment of the population in the U.S. and other western countries.

FUTURE PLANS

Prediction Sciences plans to apply for Phase II funding for this project. To that end, we have lined up Phase III support from Quest Diagnostics, Pyrosequencing and Pequot Capital. We are currently in negotiations with all three, and are pursuing Specialty Laboratories as a Phase III partner as well.

Prediction Sciences has reached agreement with the University of California, San Diego to be a part of our Phase II study to provide archived samples as well as initiate a small Clinical study of our methodology. In particular, we have retained as a consultant on Phase II Dr. Daniel O'Connor, one of the world's pre-emanate scientists in the field of hypertension. Dr. O' Connor currently is studying the pathogenesis of hypertension, as well as diagnostic autonomic drug responses, so Prediction Sciences' focus on therapeutic responses would nicely compliment this ongoing NIH-funded work.

Prediction Sciences is committed to making personalized medicine a reality. Because of the promising results of our hypertension study, we are independently funding further investigation to strengthen our model in the interim between Phase I and Phase II funding. We have negotiated to obtain 100 more samples from Genomics Collaborative and will expand our SNP genotyping to include thirty additional SNP's. Building upon our results immediately will help quicken the path to commercialization.

FINAL SUMMARY

The development of neural network processing towards the response/non-response outcome was successfully developed and tested on existing NIH published gene expression data.

Despite the obstacles, our preliminary results demonstrate great promise in the use of patient-specific variables to predict response to medication on an individual basis, even with a small data training set. Chi Squared analysis evaluated correlation, and revealed the SNP's and patient characteristics most correlated with ACE-Inhibitor response. Upon neural net training, this translated to 83% accuracy of the predicted population, better than 40-50% accuracy which is the current state-of-the-art.

These results provide strong evidence that pharmacogenomic diagnostics for hypertension will:

1) Require a multigenic approach
2) Require the inclusion of patient medical information
3) Require the usage of a neural network interpolative algorithm to best predict response

Overall, Prediction Sciences believes this study points the way to improving patient care, reducing medical expenditures, and increasing the efficiency of modern medicine.

Appendix A

Documentation for Perl Data Processing
Dr. Scott Arouh
Prediction Sciences HT-SBIR
August - October, 2001

Introduction

A Perl script performs preliminary data analysis and house-keeping functions on a given genotype data file. It is anticipated that it will be used in conjunction with a Perl programmer, as it may be desired to run successive major steps in its main program only after observation and modification of the output by a human observer. The script is called "genotype_process.pl"; it is found in the directory "...\Drug Response Predictor\perl\Genotype_Processing\scripts\", and its three major processing steps are clearly indicated (and easily turned off) in its main code block.

Input Genotype Data File

The given genotype data file contains all the input data that will be used in construction of the response predictor. Its assumed format is a tab-delimited text file with the first row containing column labels, and with each succeeding row containing field values for a different patient. One of the columns is assumed to be "Patient #", and one is assumed to be "Response". The field values are each assumed to be simple binary values (0 if the genotypic feature is not present in the patient, 1 if it is). This data file is specified within the script as $raw_input_fnam (e.g. "...\Drug Response Predictor\perl\Genotype_Processing\data\raw_input\Test_Genotype_Data_8_25_01.txt").

Preliminary Preprocessing to be Done by Hand: Before running this script, the user must assemble the input genotype data file. This includes the user choosing which of the genotypic variables are to be included in the data file.

In the original genotype data file, we should probably include, for each SNP of e.g. C/T, the possibilities CC, CT, TT, C, and T. In this example, C refers to "at least 1 C": e.g. "C" = CC or CT and "T" = TT or CT. Whether a given column is used in generating predictions should be determined with respect to the fractions f_NR and f_R, the fractions of patients having the characteristic under consideration (a "1" in the current column) that are also non-responders and responders (according to the given correct output field), respectively. The numbers of data points (patients) contributing to these calculations, N_NR and N_R, are also relevant. In principle, we should then keep only columns corresponding to at least one of the following two cases: (1) largest f_NR/f_R >= thresh that still has N_NR+N_R >= min; (2) largest N_NR+N_R >= min that still has

f_NR/f_R >= thresh. That is, we should keep 0, 1, or 2 of the 5 possible combinations listed above for each putative SNP.

This choice of data columns to use in generating predictions is to be done by hand. This is the most appropriate technique at this stage for a variety of reasons. First, this is a research phase and we do not know whether we will have too many or too few useful columns of data. Second, this method allows us to avoid restricting the format of the genotype data file too much. Finally, data from both chromosomes may not even be present, in which case only the C and T columns above would be available and any automated technique we implement will need significant modification.

Three Primary Functions

There are three primary functions of the genotype_process.pl routine. Step 1: it categorizes the SNPs in the genotype data file. Step 2: it applies crude tests to each of these SNP groups. Step 3: it pools the results of the crude tests for each of the SNP groups and applies a crude test to these results. The goal of each crude test, whether applied to a single SNP group or to results of previous tests, is to predict the patient response.

Each of these three primary functions may be replaced by human or another program. In Step 1, the automatic categorization of SNPs, an observer may wish to alter the groups of SNPs (subject to the constraint that each SNP appear in only a single SNP group). In Step 2, in which crude predictors of patient response are constructed for each SNP group, an observer may wish to replace the results of these crude tests with those of neural networks. In Step 3, in which the results of preliminary tests are pooled, a neural network may also be used to replace the crude predictors.

Crude Predictors

The default versions of the predictors implemented in Steps 2 and 3 are what we call "crude predictors." These predictors take the form of a function "AtLeastK." As the name implies, AtLeastK(k) returns true if at least k of its inputs are true, and false otherwise. If only "problematic" SNPs are included in the genotype, i.e., those correlated with non-response, and if the "Response" field given in the genotype data file really measures non-response to a given therapy, then we intuitively expect more problematic SNPs to correlate more strongly with the non-response output. We therefore intuitively expect that the AtLeastK(k) function, when called with higher value of k, should become a more accurate predictor of non-response. The problem with this approach is amount of data: the number of patients with at least k problematic SNPs is likely to decrease quickly with increasing k. As a practical matter, we suspect we will only get reproducible statistics with k equal to 2 or 3; we therefore only evaluate AtLeastK(k) for k={1,2,3,4}.

We choose these values for the AtLeastK function based on an analysis of the expected probabilities of detection an of false alarm. We determined from a patent granted to Pyrosequencing that several SNP's in the ACE inhibitor system appeared with frequencies of the order of 5% in normal patient populations but 15% in diseased populations. Based on the assumption that there were 3 independent groups of 10 SNPs, each of which showed the same preferential presence in the diseased population found as listed above, a binomial distribution argument indicates that the probability a crude

$$P(T^{(k;N)}) = 1 - \sum_{i=0}^{k-1} \binom{N}{i} p^i q^{N-i}$$

predictor AtLeastK(k=2; N=10) is as follows:

$$P(T^{(k;N)})(p)\Big|_{k=2,N=10} = \begin{cases} 0.086 & \text{(responder)} \\ 0.77 & \text{(nonresponder)} \end{cases}$$

Here, we have relabeled AtLeastK(k;N) with T(k;N). Furthermore, pooling N=3 such tests together with another crude predictor, the same formula gives

$$P(U^{(k;N)})(p)\Big|_{k=2,N=3} = \begin{cases} 0.021 & \text{(responder)} \\ 0.87 & \text{(nonresponder)} \end{cases}$$

Here, we label the AtLeastK(k;N) function with U(k;N) to emphasize the different context (the different input space) of this function. Combining this pooled test with its negation,

$$P(\overline{U}^{(k;N)})(p)\Big|_{k=2,N=3} = \begin{cases} 0.98 & \text{(responder)} \\ 0.13 & \text{(nonresponder)} \end{cases}$$

we find that (assuming a prior probability of 50% responder, 50% nonresponder for a given patient) the pair of tests U and NOT(U) will issue predictions to about 93% of all patients, and these predictions will be correct about 93% of the time. (These two values of 93% are only coincidentally nearly equal in this case.)

Such performance measures would be marketable if they could be realized. Crude predictors are therefore an appropriate place to start our construction of a drug response predictor, as we will have a benchmark against which to compare our numerical results.

Output Files

An observer assesses the results of the analysis performed by genotype_process.pl through several files, all appearing in the directory "...\Drug Response Predictor\perl\Genotype_Processing\data\processed\". We describe the output files from each of the three major processing steps in succession.

The output for the first of these steps, Step 1 (SNP categorization), is a set of SNP groups described by the file named $flag_fnam ("SNP_Group_flags.txt" by default); it contains two columns: flags indicating (non-negative) group number, and labels of corresponding columns taken from the original genotype data file. The group numbers provided here may be changed to any other non-negative number: a non-negative number here defines a SNP group; multiple SNPs having the same such number will be grouped together. If a SNP is to be ignored, set its value to -3.

The output for the second of these steps, Step 2 (preliminary testing), is a series of files, one for each SNP group, containing the output of the preliminary tests. These files are contained in the directory "SNP_Group-Specific" (which, we remind the reader, is still within the "processed" directory, as are all of the data files in this discussion). Each SNP group gets two output files in this step. The first of these files, with a name such as "SNP_Group_0_Crude_Test_Results.log", contains a tab-delimited table, with rows corresponding to different patients and columns to patient ID, Response, and preliminary test result. These column labels appear in the first line of the data file. The second of these data files, with a name such as "SNP_Group_0.log", contains flags indicating how each column from the original genotype data file was used in generating the preliminary results. This flags file thus contains one line for each column in the original genotype data file, with flag values of 1 indicating the corresponding field was used as an input to the tests, 2 indicating an output, 0 indicating an ignored field, and -1 indicating a patient ID field.

The output for the third of these steps, Step 3 (pooling of preliminary test results), is a quartet of files. The first, "pooled_data.dat", pools the results of all the preliminary tests into a single data file. The second, "pool_flags.txt", indicates which columns from the table in the formerly named data file were used as inputs to a pool-based predictor. The third output file, "pool_results.dat", contains a table of pool test results for each patient. The final output file, "pool_results_stats.txt", contains statistics on the degree to which the predicted outputs agree with the nominal outputs. These statistics are described in the following section.

Expected Mode of Operation

The anticipated protocol with which this routine will be used consists of the user running each of the above three steps in succession.

First, the user uses a spreadsheet program to create columns for e.g. C and T above for each SNP. This script then calculates the statistics

{f_ne, f_eq, N_ne, N_eq, f_eq/f_ne, N_eq+N_ne, valid}

and appends the results to the label of each flag file. Here, the quantities {f_eq, N_eq} correspond to the {fraction, number} of unity-valued inputs that equal the given correct output. Above, these were called {f_NR, N_NR}; the notation is changed here to allow for the fact that the output does not necessarily equal "non-response," but could instead equal "response." Furthermore, the quantities {f_ne, N_ne} correspond to the {fraction, number} of unity-valued inputs that do not equal the nominal output; these were previously called {f_R, N_R}. Note that N_eq+N_ne = N_1 (or, more descriptively, N_unity_inputs), the total number of 1's that appear in the corresponding column. Finally, the "valid" parameter is given by

valid = (f_eq/f_ne >= thresh && N_eq>=min && N_1>=min).

for some significance thresholds thresh (e.g. 1.5) and min (e.g. 10). The user should change these thresholds (hardwired in the subroutine &binary_stats()) as appropriate.

The user considers these calculated values in manually editing the default usage flag values in flags files. The user may take advantage of the "valid" parameter as calculated above, and set a flag to ignore a column if its statistics (as summarized by the "valid" parameter) indicate that it does not statistically significantly correlate with the nominal output. This is automatically implemented for the SNP grouping performed in Step 1, so that "valid=0" causes a column to be ignored in further processing. For later steps, the user must decide for himself whether to ignore a set of test results even if they produce a "valid=0" statistical measure.

The user also observes the model outputs from each processing step. One goal is for the fraction f_corr of predicted outputs (whether 0 or 1) that are correct to approach 0.9. Another is for the number N_corr of predicted outputs to approach ~0.9 * Ndata, where Ndata is the number of data points we have.

Finally, the user repeats this entire training process using a nominal output in the input genotype data file of one minus that used above. The logic here is that, in constructing a detector of e.g. the non-responder phenotype, we have focused on observables that correlate to non-response. However, we also wish to detect responders. Although we can use our non-responder detector to do this, we have ignored (combinations of) observables that correlate with the responder phenotype. We outline in the next section more specifically how we expect this combination of tests to improve our classification scheme.

Estimates of Anticipated Performance

Preliminary calculations indicate that tests constructed according to the above-outlined plan would yield decent distinguishability between the normal and diseased populations (e.g. 10% of one vs. 80% of the other would pass a given test). As discussed in the previous section, we suspect, based on our method of constructing the tests, that a corresponding test, constructed to predict one minus the output predicted with the first test, will yield improved classification ability. We now consider some actual numerical probabilities we may hope to observe.

Author's note: the following discussion is esoteric and will be replaced by actual performance measures when we have a real data set. I therefore skip the step of formatting the equations cleanly.

Consider a pair of tests T1, T2, with probabilities of detection for normal (N) and diseased (D) patient populations given as follows:

T1: P(T1|N) = 0.1
P(T1|D) = 0.8
T2: P(T2|N) = 0.8
P(T2|D) = 0.1

These tests will have corresponding prior probabilities of occurring of

P(T1) = P(T1|N) P(N) + P(T1|D) P(D) = 0.1 0.5 + 0.8 0.5 = 0.45
P(T2) = P(T2|N) P(N) + P(T2|D) P(D) = 0.1 0.8 + 0.1 0.5 = 0.45

We first consider what we can infer if only test T1 is performed. We then examine the corresponding inferences if both of these tests are performed.

With only a single test, T1, the probabilities of a normal (N) or diseased (D) patient passing or failing the test are as follows:

P(N|T1) = P(T1|N) P(N) / P(T1) = 0.1 0.5 / 0.45 = 0.11
P(D|T1) = P(T1|D) P(D) / P(T1) = 0.8 0.5 / 0.45 = 0.89
P(N|!T1) = P(!T1|N) P(N) / P(!T1) = (1-0.1) 0.5 / (1-0.45) = 0.82
P(D|!T1) = P(!T1|D) P(D) / P(!T1) = (1-0.8) 0.5 / (1-0.45) = 0.18

where '!' indicates logical NOT.

For test T1 alone, then:

If a patient passes T1, they have 89%/11% probabilities of being diseased/normal.
If a patient does not pass T1, they have 82%/18% probabilities of being normal/diseased.

The thing to note here is that a split of 89-11 is borderline useful, but 82-18 is not quite marketable. We will see this problem resolved when both tests are considered simultaneously.

Now consider the case where both tests are performed. There are four possible outcomes from such a pair of tests:

A =def !T1 & !T2
B =def T1 & !T2
C =def !T1 & T2
D =def T1 & T2

We calculate conditional probabilities for these four cases, for normal (N) and diseased (D) patient subpopulations, based on an incorrect assumption that the two tests will be independent:

P(A|N) = P(!T1 & !T2|N) = P(!T1|N) P(!T2|N) = (1-P(T1|N)) (1-P(T2|N)) = (1-0.1)(1-0.8) = 0.18

P(A|D) = P(!T1 & !T2|D) = P(!T1|D) P(!T2|D) = (1-P(T1|D)) (1-P(T2|D)) = (1-0.8)(1-0.1) = 0.18

P(B|N) = P(T1 & !T2|N) = P(T1|N) P(!T2|N) = P(T1|N) (1-P(T2|N)) = 0.1(1-0.8) = 0.02

P(B|D) = P(T1 & !T2|D) = P(T1|D) P(!T2|D) = P(T1|D) (1-P(T2|D)) = 0.8(1-0.1) = 0.72

P(C|N) = P(!T1 & T2|N) = P(!T1|N) P(T2|N) = (1-P(T1|N)) P(T2|N) = (1-0.1)0.8 = 0.72

P(C|D) = P(!T1 & T2|D) = P(!T1|D) P(T2|D) = (1-P(T1|D)) P(T2|D) = (1-0.8)0.1 = 0.02

P(D|N) = P(T1 & T2|N) = P(T1|N) P(T2|N) = 0.1 0.8 = 0.08

P(D|D) = P(T1 & T2|D) = P(T1|D) P(T2|D) = 0.8 0.1 = 0.08

The corresponding prior probabilities of these four cases are thus:

P(A) = P(A|N)P(N) + P(A|D)P(D) = 0.18
P(B) = P(B|N)P(N) + P(E|D)P(D) = 0.37
P(C) = P(C|N)P(N) + P(C|D)P(D) = 0.37
P(D) = P(D|N)P(N) + P(D|D)P(D) = 0.08

where we note that indeed, these priors satisfy P(A)+P(B)+P(C)+P(D)=1

Note that we have above treated T1 and T2 as independent only within a given population, e.g. only within N, or only within D. We find that T1 and T2 are (apparently) not independent over the whole patient population (i.e., P(T1|T2) is not equal to P(T1), even though I can't directly show this). This follows from calculations of P(A..D) based directly on calculations of P(T1) (and only secondarily on P(T1|N)) rather than directly on P(T1|N):

P(A =def !T1 & !T2) = P(!T1) P(!T2) = (1-P(T1)) (1-P(T2)) = (1-0.45) (1-0.45) = 0.30

P(B =def T1 & !T2) = P(T1) (1-P(T2)) = 0.45 (1-0.45) = 0.25

P(C =def !T1 & T2) = (1-P(T1)) P(T2) = (1-0.45) 0.45 = 0.25

P(D =def T1 & T2) = P(T1) P(T2) = 0.45 0.45 = 0.20

Although these at least satisfy P(A)+P(B)+P(C)+P(D)=1, they disagree with the previous calculations of P(A..D), which are based directly on the underlying

conditional probabilities (rather than averages thereof). The calculations here are in a sense "prior" estimates, where P(T1) are prior probabilities; the previous calculations were then "posterior" estimates, where the priors P(T1) were not used.

We calculate from the above prior and conditional probabilities the posterior probabilities that a given patient is normal (N) or diseased (D), given a given pair of test results:

A:
P(A) = 0.18
P(N|A) = P(A|N) P(N) / P(A) = 0.18 0.5 / 0.18 = 0.5
P(D|A) = P(A|D) P(D) / P(A) = 0.18 0.5 / 0.18 = 0.5

B:
P(B) = 0.37
P(N|B) = P(B|N) P(N) / P(B) = 0.02 0.5 / 0.37 = 0.027
P(D|B) = P(B|D) P(D) / P(B) = 0.72 0.5 / 0.37 = 0.973

C:
P(C) = 0.37
P(N|C) = P(C|N) P(N) / P(C) = 0.72 0.5 / 0.37 = 0.973
P(D|C) = 1-P(N|C) = 0.027

D:
P(D) = 0.08
P(N|D) = P(D|N) P(N) / P(D) = 0.08 0.5 / 0.08 = 0.5
P(Dis|D) = 1-P(N|D) = 0.5

We see that for cases A and D, which occur 26% (=18%+8%) of the time, we won't be able to say anything about the disease state of the patient. However, for the remaining 74% of the patients, we will be able to say with 97% certainty what their disease state is.

However, with only a single test, we found above that P({N,D}|!T1)={0.82,0.18}, which distinguishes normal from disease in this case much better than the composite case of A =def (!T1 & !T2). Now it becomes important that tests T1 and T2 will not, in fact, be independent for the data set: patients will actually either be normal or diseased, so their test results for T1 and T2 should be related. This relation will manifest itself in the form of tests T1 and T2 being exclusive to some extent: for example, we expect that P(T1 & !T2) should be greater than (the value of 0.37) calculated above, and P(!T1 & !T2) should be less than that calculated above (0.18).

We thus anticipate, based on an incorrect assumption of independence, that the pair {T1,T2} of tests will yield dramatically increased distinguishability most (74%) of the time, but dramatically decreased distinguishability the remainder of

the time. We also anticipate, however, that correcting for the dependence of these tests would only increase the fraction of the time that the high distinguishability predictions would be made. We thus anticipate that the pair of tests would significantly improve performance. Once the tests are constructed and run with real data, the priors P(A..D) and the posterior probabilities P({N,D}|{A..D}) will need to be calculated: these will be the deliverable quantities directly useable in the clinic.

The upshot of the previous discussion is that a list of probabilities will characterize the performance of our crude tests. If we construct a pair of tests T1, T2, for which we measure the probabilities of detection for normal (N) and diseased (D) patient populations, then we will need to separately measure the probabilities of each possible outcome of these tests (i.e., T1 predicts N and T2 predicts N) in each patient population (i.e., N or D). These probabilities will be he deliverable quantities directly useable in the clinic, since a given patient's genotype will produce predictions of either N or D for each test (T1 or T2).

Appendix B

MATLAB NEURAL NETWORK PROCESSING OF DRUG RESPONSE DATA

A Matlab script performs a variety of neural network analyses on a variety of processed data files generated from genotype_process.pl. It is anticipated that it will be used in conjunction with a Matlab programmer, as it may be desired to run successive major steps in its main program only after observation and modification of the output by a human observer. The script is called " DrugResponse_Processing_Session.m"; it is found in the directory "...\Drug Response Predictor\ Matlab_Neural_Net_Processing\M-Files\", and its major processing steps are clearly indicated in both its main code block and in its output (to Matlab's standard output).

There are two tasks to perform. (1) Neural net processing of data for each SNP group. (2) Neural net processing of pooled data of all SNP groups. The majority of our work has focused on performing the housekeeping functions of deriving neural network inputs from and storing the network outputs to appropriate files. We therefore summarize these activities in detail in following sections. Before we delve into this listing of file names, we first review the techniques we use to train and test any of our neural networks once we have assembled inputs from appropriate data files.

Training Neural Networks in Matlab

To perform the actual training and testing of the neural networks, we employ Matlab's Neural Network Toolbox. The M-files used reside in the directory "...`\Matlab_Neural_Net_Processing\M-File`" and all of the subdirectories of "...`\Matlab_Neural_Net_Processing\M-File\Utilities`".

The M-file "...`\Utilities\Neural_Net_Routines\train_nn.m`" contains our neural network training function. We have experimented with several neural network learning rules and programmer-supplied architectures. These learning rules all apply to fully connected feed-forward mapping networks, the architecture employed by industry-standard backpropagation networks. Although the learning rule we used is that of scaled conjugate gradient descent, the only qualitative difference among the learning rules should be the speed of training. This speed is essentially irrelevant to this project, as programmer time vastly outweighs training times of 10 or 20 seconds. The other item mentioned above was support for programmer-supplied architectures. This refers to the number of layers and neurons in each layer that make up the neural network. The network we trained contains a single hidden layer of 20 neurons, and an output layer containing a lone neuron corresponding to the single output from this system.

"`train_nn.m`" also contains the data pre- and post-processing steps performed. These consist of a data grouping step, a normalization step and a principal component analysis step. The **data grouping** step consists of breaking the training data up into training, validation, and testing data. We choose these data subsets to consist of 50%, 25%, and

25% of the full given data set, and we draw them in an interleaved process so each subset contains data from all portions of the data set. The net is trained to the training data; training is stopped when testing the net on the validation data begins to yield an increasing rather than a decreasing error; and the net performance is characterized with the testing data set. The **normalization** step normalizes each component of the full given data set (including training, validation, and testing, and including the outputs as well as the inputs) to have zero mean and unity standard deviation. The **principal component analysis** step consists of obtaining the principal components (linear superpositions of inputs) of the full data set and ignoring those that contribute less than 1% of the variance in this (already normalized) data set. This has the effect of reducing the number of inputs to the network from e.g. 9 to 6 for a preliminary training session.

Training stops at 100 epochs or when testing error begins to significantly increase, whichever comes sooner. The default architecture contains a single hidden layer of 40 neurons and a single output neuron corresponding to the single network output.

To analyze the performance of each network we train, we perform a regression analysis on the network outputs. This consists of plotting the predicted output vs. the actual output, fitting with a best-fit line (y=mx+b), and calculating the correlation coefficient, R. This analysis must be performed separately on the training and testing data. Optimal values for the best-fit line parameters are m=1 and b=0. The extent to which this best-fit line is not ideal is measured by the correlation coefficient, R, which is loosely a measure of the fraction of the variance of the data accounted for by the model: an R approaching 1 is ideal. Statistics of the percentage error of the predictions (mean, standard deviation, and root-mean-square percent error) are also calculated and presented. To insure that the model embodies some nontrivial predictor, we also calculate these error measures for a null prediction consisting of a constant value, the mean of the output values over the training data set. In cases where the user is not confident, based on the best-fit line and corresponding correlation coefficient calculation, that the network is performing adequately, he may derive some confidence in the performance of the network if the mean percent error from the network is significantly smaller than that for the null predictor.

NEURAL NETWORK PROCESSING OF DATA FOR EACH SNP/VARIABLE

Input files: (1) A raw input data file, e.g. "...\data\raw_input\Test_Genotype_Data_8_25_01.txt", with a format as follows: column labels (which get ignore here) in the first line; remaining lines are tab-delimited values. Fields to use are as specified in flags files. (2) A file listing files containing usage flags: "...\data\processed\usage_flags_fnam.log". Format: one file name per line, including path. (3) Flags files listed in (2), each with the following format: tab-delimited lines, with the first column containing the usage for the corresponding column in the raw data file (1). These flag values have meanings as summarized in the following table:

Flag Value	Meaning
-1	Patient ID (Ignore in neural net)
0	Ignore
1	Input to neural net
2	Output to neural net

Output files: (1) Tables of results, with file names based on the flag file names (listed in usage_flags_fnams_fnam, including the path, if any), but with '_Neural_Net_Results' appended just before file name extension. For example, for a flags file "...\SNP_group_0.log", the output file name would be "...\SNP_group_0_Neural_Net_Results.log'. The format of these results files is as follows: tab-delimited columns, including "Patient_ID," "Response" (actual, given output), and "Net_Response" (output of neural net). The first line contains column labels; succeeding lines contain values. These results must be interpreted with respect to how each data point was used during training: evenly distributed portions of the given data are used for training (50%), validation (25%), and testing (25%). (2) Trained network data file: after each network training, a net_with_params object (described below) is saved to a MAT file named "...\Matlab_Neural_Net_Processing\data\$NAME_Trained_Net.mat", where $NAME is the base file name (not including path or extension) of the flags file used to assemble the data for the current network.

Neural Network Processing of Pooled Data of All SNP Groups.

This step is performed twice: once using the predicted outputs generated above (from separate neural networks for each SNP groups), and a second time using previously compiled crude predictor results for each SNP group.

For the pooling of network outputs, there are no additional input files and two output files. The output file name for predicted network output is based on the path of the first flags_fnam used above to train individual SNP group nets and the base file name of 'pooled_net_results.dat'. The trained network structure is stored to the same path as the trained networks above, but with name 'pooled_net_results_Trained_Net.mat'.

For the pooling of crude predictors, the file names are as follows:
Input files: (1) A single usage flags data file ("...\data\processed\pool_flags.txt"), with a format identical to flags files described above. (2) A data file ("...\data\processed\pooled_data.dat"), with a format identical to the raw data file described above.
Output files: (1) Tables of results, with file names based on the flag file names (listed in usage_flags_fnams_fnam, including their path, if any), but with '_Neural_Net_Results' appended just before file name extension. The format of these files is identical to the output file described above. These files have the same distribution of data used for

training, validation, and testing. (2) The trained neural network is stored in the same fashion as above.

obtained from the New York State Department of Health were extracted and tested for the presence of WNV-NY1999 sequences by real-time 5′ nuclease PCR. Five of the serologically confirmed cases and none of the controls were positive for WNV-NY1999 sequences (NS3, 3/5 patients; NS5, 5/5 patients; samples obtained 9–16 days after onset of illness). Although it was not possible to analyse matched sets of CSF and sera for all patients, one of the individuals negative by real-time PCR of CSF was positive by PCR for NS5 sequences in serum.

The establishment of a real-time PCR method for detection of WNV sequences in human CSF improves diagnosis of viral encephalitides. Although CSF containing WNVs other than WNV-NY1999 were not available for analysis, the primer/probe sets described here are predicted to detect lineage I WNVs, viruses associated with outbreaks of acute illness. Our results suggest that the detection of WNV-NY1999 sequences in CSF correlates with a poor prognosis particularly in older individuals. Further investigation is needed to find whether this correlation can be extended to other flavivirus encephalitides. As antiviral research identifies drugs with activity against WNV, the ability to rapidly implicate this virus is anticipated to achieve clinical importance similar to that associated with herpesviral or enteroviral diseases.

We are grateful to Leo Grady, Cinnia Huang, and Susan Wong of the New York State Department of Health for providing CSF and serum samples, Rob Lanciotti and John Roehrig from the Division of Vector-Borne Infectious Diseases, CDC, for WNV extracts, and clinical and serological results, and Charles Calisher and Ingo Jordan for helpful discussions.

1 Jeffery KJM, Read SJ, Peto TEA, Mayon-White RT, Bangham RM. Diagnosis of viral infections of the central nervous system: clinical interpretation of PCR results. *Lancet* 1997; **349**: 313–17.

2 Asnis D, Conetta R, Waldmon G, et al. Outbreak of West Nile-like viral encephalitis: New York 1999. *MMWR Morb Mortal Wkly Rep* 1999; **48**: 845–49.

3 Higuchi R, Fockler C, Dollinger G, Watson R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biol/Technology* 1993; **11**: 1026–30.

4 Lee LG, Connell CR, Bloch W. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res* 1993; **21**: 3761–66.

5 CDC. Case definitions for infectious conditions under public health surveillance. *MMWR Morb Mortal Wkly Rep* 1997; **46**: 1–55.

Emerging Diseases Laboratory, Departments of Neurology, Microbiology and Molecular Genetics, Anatomy, and Neurobiology, University of California, Irvine, CA 92697-4292, USA (T Briese PhD, W G Glass BS, Prof W I Lipkin MD)

Correspondence to: Dr T Briese
(e-mail: tbriese@ucl.edu)

Pharmacogenetic prediction of clozapine response

M J Arranz, J Munro, J Birkett, A Bolonna, D Mancama, M Sodhi, K P Lesch, J F W Meyer, P Sham, D A Collier, R M Murray, R W Kerwin

We did association studies in multiple candidate genes to find the combination of polymorphisms that give the best predictive value of response to clozapine in schizophrenic patients. A combination of six polymorphisms in neurotransmitter-receptor-related genes resulted in 76·7% success in the prediction of clozapine response (p=0·0001) and a sensitivity of 95% (±0·04) for satisfactory response. These results will form the basis for a simple test to enhance the usefulness of clozapine in psychiatric treatment.

The atypical antipsychotic clozapine was reintroduced into the UK and USA in 1990, following the demonstration of superior efficacy and tolerability in severely treatment-resistant patients.[1] Since then a range of well-tolerated atypical antipsychotics have been introduced.[2] Although the evidence base for the use of these drugs is compelling, various reasons seem to prevent use in greater numbers of eligible patients. The treatment costs are greater than with classic antipsychotics and the response to these drugs is heterogeneous, with between 30% and 60% responding to clozapine, the archetypal atypical antipsychotic.

We have attempted to explore new targets by studying pharmacogenetic associations in a large number of patients treated with clozapine. We have previously shown that allelic variation in the serotonin neurotransmitter receptor 2A gene (*5-HT$_{2A}$*) is a factor in determining clinical response to clozapine.[3,4] However, *5-HT$_{2A}$* polymorphisms on their own cannot fully explain the variability seen in treatment response, and it has been postulated that there are contributions from other mutations in neurotransmitter-receptor-related genes.

We did association studies in multiple candidate genes to find the combination of polymorphisms that give the best

Gene	Polymorphism	Detection method (restriction enzyme/electrophoresis conditions)
Adrenergic receptors		
α_{2A}	−1291-C/G	*Mspl*/10% PAGE
α_{2B}	−261-G/A	*Hhal*/10% PAGE
α_{1A}	Arg492Cys	*Pstl*/3% agarose
Dopamine receptor		
D_3	Ser9Gly	*Mscl*/3% agarose
Serotonin receptors		
5-HT$_{2A}$	−1438-G/A	*Mspl*/2% agarose
5-HT$_{2A}$	102-T/C	*Mspl*/2% agarose
5-HT$_{2A}$	516-C/T	*Sau96I*/2% agarose
5-HT$_{2A}$	His452Tyr	*Bbvl*/3% agarose
5-HT$_{2A}$	Thr25Asp	*BstNI*/2% agarose
5-HT$_{2C}$	Cys23Ser	*Hinfl*/4% agarose
5-HT$_{2C}$	−330-GT/−244-CT repeat	10% PAGE
5-HT$_{3A}$	178-C/T	10% PAGE/5% glycerol
5-HT$_{3A}$	1596-G/A	*Nhel*/3% agarose
5-HT$_{5A}$	12-A/T	*Bsrl*/2% agarose
5-HT$_{5A}$	−19-G/C	*BsaJI*/10% PAGE
Serotonin transporters		
Transporter *5-HTT*	VNTR	4% agarose
Transporter promoter *5-HTT*	5-HTTLPR	4% agarose
Histamine		
H_1	Leu449Ser	*Bsml*/10% PAGE
H_2	−1018-G/A	*HaeIII*/10% PAGE

PAGE=Polyacrylamidegel electrophoresis.

Table 1: **List of polymorphisms studied and detection methods for the identification of the alleles**

Polymorphism	p for genotype	p for allele
−1291-C/G (ADR2A)	0·85	0·61
−261-G/A (ADR2A)	0·54	0·40
Arg492Cys (ADR1A)	0·22	0·10
Ser9Gly (D3)	0·89	0·63
Leu449Ser (H1)	0·30	0·35
−1010-G/A (H2)	0·08	0·43
His452Tyr (5-HT2A)	0·01*	0·02
Thr25ASP (5-HT2A)	0·57	0·78
−1438-G/A (5-HT2A)	<0·001	0·001
102-T/C (5-HT2A)	<0·001	0·001
516-C/T (5-HT2A)	0·65	0·82
−330-GT/−244-CT (5-HT2C)	0·04†	0·31
Cys23Ser (5-HT2C)	0·08‡	0·17
178-C/T (5-HT3A)	0·92	0·79
1596-G/A (5-HT3A)	0·98	0·85
−12-A/T (5-HT5A)	0·20	0·09
−19-G/C (5-HT5A)	0·55	0·35
5-HTTLPR	0·04	0·36
VNTR (5-HTT)	0·89	0·70

*Try recessive: 0·004. †Short allele dominant. ‡*Ser23* dominant.

Table 2: **Comparison of responders vs non-responders for the 19 polymorphisms genotyped**

predictive value of response. 19 genetic polymorphisms (16 previously described and three new genetic variants) in nine clozapine-targeted receptor subtypes and a neurotransmitter transporter were studied in a sample of 200 schizophrenic patients treated with the drug (table 1). Treatment response was retrospectively assessed using the global assessment scale.[5] 133 patients were classified as responders and 67 as non-responders. All patients were white Caucasians of British origin. Informed consent was obtained from all the patients included in the study.

Table 2 summarises the results. Comparisons of allele and genotype frequencies between responders and non-responders were done for each polymorphism using χ^2 or Fisher tests. Odds ratios with 95% Cornfield CIs were also calculated for allele frequencies. The prediction levels were calculated by logistic-regression analysis using response to clozapine as the dependent variable and the polymorphisms studied as independent variables. As the 5-HT_{2A} polymorphisms 102-T/C and −1438-G/A were found in almost complete linkage disequilibrium, only one of them (102-T/C) was included in the regression analyses. A combination of the six polymorphisms showing the strongest association with response ($p<0.09$; 5-HT_{2A} 102-T/C and His452Tyr, 5-HT_{2C} −330-GT/−244-CT and Cys23Ser, 5-HTTLPR, H2 −1018-G/A) gave a level of prediction of 76·86% ($\chi^2=35.8$; $p=0.0001$). Although the 5-HT_{2C} gene is located in the X chromosome, the short form of the −330-GT/−244-CT polymorphism and the Ser23 allele were considered dominant, obviating the need to calculate separate analysis for men and women. The positive predictive value for these results was 0·76 (SD 0·08), and the negative predictive value was 0·82 (0·16). This combination had a sensitivity of 95·89 (0·04) for the identification of patients who improved satisfactorily with treatment and a specificity of 38·3 (0·14) for the detection of individuals who did not show a substantial improvement in response to clozapine treatment. Simpler combinations of genotypes could also be useful for the determination of the individual's response. Possession of both genotypes, T102/– and His452/His452 in the 5-HT_{2A} receptor, was associated with good response to clozapine in 80% of the patients. Although only about 50% of the patients showed this genotype combination, this could prove a simple method of identifying individuals likely to benefit from clozapine treatment.

These results constitute the first report of the use of pharmacogenetics for the individualisation of psychiatric treatment. Predictability testing is one of the main aims of pharmacogenetics and this report represents an early realisation of this potential. If our results are prospectively validated they will form the basis for a simple test to enhance the usefulness of this expensive drug in a heterogeneously responsive group of patients and also provide a test to determine the benefit of persevering with treatment in poor responders. This test will have implications for the cost-effective and rotational prescribing of clozapine. In addition, more patients will benefit from clozapine treatment if a positive response is predicted. The findings of this study highlight the potential of pharmacogenomic studies as the key for the future improvement and individualisation of clinical treatment.

M J Arranz was supported by the Diana fellowship.

1 Kane J, Honigfeld G, Singer J, Meltzer H. Clozapine for the treatment-resistant schizophrenia: a double-blind comparison with chlorpromazine. *Arch Gen Psychiatr* 1988; 45: 789–96.

2 Kerwin RW, Taylor D. New antipsychotics, a review of their current status and clinical potential. *CNS Drugs* 1996; 6: 71–82.

3 Arranz MJ, Collier D, Sodhi M, et al. Association between clozapine response and allelic variation in 5-HT_{2A} receptor gene. *Lancet* 1995; 346: 281–82.

4 Arranz MJ, Munro J, Sham P, et al. Meta-analysis of studies on genetic variation in 5-HT_{2A} receptors and clozapine response. *Schiz Res* 1998; 32: 93–99.

5 Endicott J, Spritzer RL, Fleiss JL, Cohen J. The global assessment scale. *Arch Gen Psychiat* 1876; 55 (suppl): S5–12.

Department of Clinical Neuropharmacology (M J Arranz PhD, J Munro MRCPsych, J Birkett BSc, A Bolonna BSc, D Mancama BSc, M Sodhi PhD, R W Kerwin FRCPath), **Department of Psychological Medicine** (P Sham MRCPsych, D A Collier PhD, R M Murray FRCPsych), **Institute of Psychiatry, Denmark Hill, London SE5 8AF, UK; and Department of Psychiatry, University of Würzburg, Würzburg D-97080, Denmark** (K P Lesch MD, J F W Meyer MD)

Correspondence to: Dr M J Arranz
(e-mail: marranzc@hgmp.mrc.ac.uk)

Methotrexate effects in patients with rheumatoid arthritis with cardiovascular comorbidity

Robert B M Landewé, Ben E E M van den Borne, Ferdinand C Breedveld, Ben A C Dijkmans

Methotrexate, an antirheumatic drug that may increase serum homocysteine, significantly increases mortality in patients with rheumatoid arthritis and cardiovascular comorbidity.

Methotrexate (MTX), the most frequently used disease-modifying antirheumatoid drug (DMARD) for rheumatoid arthritis, might decrease mortality in rheumatoid arthritis.[1] Mortality is thought to be increased because of an excess of cardiovascular deaths. Concentrations of homocysteine, which may promote atherosclerosis and thrombosis,[2] are commonly increased in patients with rheumatoid arthritis.[3] Haagsma and colleagues have shown that long-term low-dose MTX treatment may further increase concentrations of homocysteine.[4] These observations make it conceivable that MTX treatment may contribute to increased mortality in patients with rheumatoid arthritis by promoting atherosclerosis.

We did a retrospective cohort study to analyse whether long-term MTX treatment was related to increased mortality in patients with rheumatoid arthritis who already had evidence of atherosclerotic vascular disease. Patients, from various rural and university hospital out-patient rheumatology clinics, were included if they had active rheumatoid arthritis, which made the start or the change of DMARD therapy necessary. The cohort had been used for other studies previously described.[5] For each patient in the cohort the start of follow-up (between September 1984 and September 1990) coincided with the start of the new DMARD (called index-DMARD). Follow-ups were censored at 1 Dec, 1995 or earlier in case of death. Detailed information about the kind and treatment duration of DMARDs was available. Patients with documented evidence for the presence of peripheral or central atherosclerotic vascular disease and/or hypertension were considered to have a cardiovascular disease profile. Only presence of cardiovascular disease at the start of follow-up was taken into consideration.

623 patients with rheumatoid arthritis (median age 56 [range 20–74] years, 71% female, median disease duration 4·5 [0·2–40] years, 82% rheumatoid-factor positive, 70% with erosive disease) were included in the cohort. The median number of DMARDs used prior to follow-up was two (range 0–7). The median duration of index-DMARD treatment was 2·2 years (3·5 years for MTX; 1·8 years for other index DMARDs).

During follow-up 73 patients died. The influence of the presence of cardiovascular disease on mortality during follow-up seemed to be greater in patients who had started MTX (crude risk ratio 4·1) than in patients who had started other

EXHIBIT C

<http://biz.yahoo.com/prnews/020425/nyth204_1.html>

They appear to be working for
AstraZeneca, Biogen, Johnson & Johnson, and Pfizer,
and either in conjunction with or against
Merck and Bristol-Myers.

They say they've found 29 haplotypes spread over 27 genes, each present in at least 10% of their patients. Of these 29, 25 had drug-specific significance, and 4 had statin-class-specific significance. They checked 100 genes spread over 3 functional pathways:
- lipid metabolism and transport,
- inflammation,
- drug metabolism.

They do not indicate how strong their correlations are. (Deal-breaker.)

The outputs they consider are effects of statin therapy:
- HDL (bad cholesterol) levels;
- LDL (good cholesterol levels);
- triglycerides (presumably, levels of).

(Their) business plan:
- License markers "to the biopharmaceutical industry for new drug indications
and life-cycle management strategies",
- (Sounds extremely vague to me; what do these terms mean?)
- Discovery of next generation drugs,
- Development of a HAP-Drug.
- Deemphasized: Diagnostic test to identify patients who may be candidates
for combination therapy or other therapeutic interventions to improve their
cholesterol status.
In other words, drug discovery, drug development (both in-house and by licensing),
and maybe diagnostic tests for combination therapy (implying there's no market
for single-therapy response prediction).

Good, real focused. [redacted] gone public and had a market capitalization
of over $320M less than a year ago with a business plan like that. Granted, that's currently down to $35.5M, but it certainly drives home the point that

the complete absence of both functioning technology and an even mildly appropriate
business plan were still amenable to funding back in 1997 (which I think is when
they were founded).

-Scott

Thursday April 25, 7:01 pm Eastern Time
Press Release
SOURCE: Genaissance Pharmaceuticals, Inc.
Genaissance Pharmaceuticals Discovers Genetic Markers Associated With Response to Individual Statin Drugs
Ongoing Analyses of STRENGTH I Study Further Demonstrate Ability of HAP(TM) Technology to Identify Drug Response Markers of Commercial Utility
NEW HAVEN, Conn., April 25 /PRNewswire-FirstCall/ -- Genaissance Pharmaceuticals, Inc. (Nasdaq: GNSC - news) today announced new results from ongoing analyses of the STRENGTH I clinical study that further demonstrate the ability of its HAP(TM) Technology to identify specific genetic markers (gene haplotypes or HAP(TM) Markers) that are associated with the effects of statin therapy, including LDL (bad) cholesterol, HDL (good) cholesterol and triglycerides. The Company believes these positive results strengthen the commercial value of its HAP(TM) Markers for the development of pharmaceutical and diagnostic products.

One hundred genes were examined so far from three functional pathways (lipid metabolism and transport, inflammation, and drug metabolism). A total of 29 HAP(TM) Markers from 27 of these genes were found to have statistically significant associations with clinical response (LDL, HDL and/or triglycerides) to simvastatin (Zocor®(1)), atorvastatin (Lipitor®(2)) or pravastatin (Pravachol®(3)). Each HAP(TM) Marker occurred in at least 10 percent of the study group and, thus, may have wide clinical applicability across the population of patients with hyperlipidemia. Twenty-five of the markers were linked to outcomes for specific drugs and four were associated with the effects of statins as a drug class. The Company believes that these important findings highlight the differences between drugs in the statin class and clearly indicate the need and the potential to optimize therapy based on the genetics of different patient populations.

``The medical community has been aware of clinical and metabolic differences among the statins, but now, for the first time, we have some genetic evidence that begins to explain these differential effects," said Antonio Gotto, M.D., Dean of the Weill Cornell Medical College and Chair of the STRENGTH Steering Committee. ``It is exciting to see the genetic

underpinnings of drug response come to light, and I expect further advances as we continue to mine this very rich collection of information."

The Company highlighted two examples of drug-specific markers using a deeper analysis that was completed for 25 genes. A marker from one gene was associated with a positive HDL cholesterol clinical response in one of the statins, a negative response in a second drug and no change in HDL response in a third. It is notable that the same marker predicts opposing effects for two of the drugs. This marker, found in 26 percent of the STRENGTH patients, illustrates how differently the drugs interact with biological pathways and may help to explain the variation in the HDL response of patients when taking different statins. In this study, this one predictive marker is as powerful as all of the other currently available predictors put together, including age, alcohol use, smoking, gender, body mass index, exercise and baseline HDL levels.

A HAP(TM) Marker from a different gene family, found in 16 percent of the study population, showed results of similar magnitude. Further analyses of STRENGTH data, which will continue over the next few months, will include the evaluation of additional markers for drug efficacy as well as markers that may be predictive of side effects, such as muscle damage.

``Through STRENGTH, we have achieved the first large-scale clinical demonstration of the ability of our HAP(TM) Technology to discriminate between these drugs," said Gualberto Ruano, M.D., Ph.D., CEO of Genaissance. ``We are extremely pleased with our findings to date which place Genaissance in an excellent position to commercialize these data by licensing markers to the biopharmaceutical industry for new drug indications and life-cycle management strategies, discovery of next generation drugs, or development of a HAP? Drug - a product that is targeted to specific patient populations. In addition, our markers could be used in a diagnostic test to identify patients who may be candidates for combination therapy or other therapeutic interventions to improve their cholesterol status."

The Company intends to present STRENGTH data at major medical meetings. The Company will be seeking patent protection for these HAP(TM) Markers and their use in pharmaceutical and diagnostic product development.

The STRENGTH (Statin Response Examined by Genetic HAP(TM) Markers) I Study, initiated in April 2001, evaluated two different doses of simvastatin (Zocor®), atorvastatin (Lipitor®) and pravastatin (Pravachol®) in over 500 patients. The population of patients in the STRENGTH Study mirrors those identified as eligible for statin therapy in the National Cholesterol Education Program (NCEP) guidelines issued last year.(4)

Genaissance Pharmaceuticals, Inc. is the world leader in the discovery and

use of human gene variation for the development of personalized medicines. The Company markets its technology and clinical development skills to the pharmaceutical industry as a complete solution for improving the development, marketing and prescribing of drugs. The Company also has identified candidates for development in its own pipeline of products utilizing its proprietary genetic markers. Genaissance has agreements with three of the top five pharmaceutical companies as well as one of the premier biopharmaceutical companies: AstraZeneca, Biogen, Johnson & Johnson and Pfizer. Genaissance is located in Science Park in New Haven, Connecticut. Please visit <http://www.genaissance.com> for additional information.

This press release contains forward-looking statements, including statements about the ability of Genaissance to apply its technologies to the development, marketing and prescribing of drugs, and the expectations as to clinical trials data. Such statements are subject to certain factors, risks and uncertainties that may cause actual results, events and performance to differ materially from those referred to in such statements, including, but not limited to, the extent to which genetic markers (haplotypes) are predictive of drug efficacy and safety, the adoption of our technologies by the pharmaceutical industry, the timing and success of clinical trials, competition from pharmaceutical, biotechnology and diagnostics companies, the strength of our intellectual property rights and those risks identified in our Annual Report on Form 10-K filed with the Securities and Exchange Commission on March 7, 2002. The forward-looking statements contained herein represent the judgment of Genaissance as of the date of this release. Genaissance disclaims any obligation to update any forward-looking statement.

1. Registered trademark of Merck & Company
2. Registered trademark of Pfizer Inc.
3. Registered trademark of Bristol-Myers Squibb Company
4. Third Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel

III) by the National Cholesterol Education Program of the National

Heart, Lung, and Blood Institute, National Institutes of Health

SOURCE: Genaissance Pharmaceuticals, Inc.

EXHIBIT D

Vol. 58 No. 1, January 2001

Letters to the Editor

PDF OF THIS ARTICLE

See Related: Authors' Articles

Return to Table of Contents

INDEX OF FIGURES AND TABLES

Genetic Determinants of Clozapine-Induced Agranulocytosis: Recent Results of HLA Subtyping in a Non-Jewish Caucasian Sample

The incidence of agranulocytosis in clozapine-treated patients is comparatively high despite the undisputed clinical advantages of clozapine.[1] The mechanisms of clozapine-induced agranulocytosis (CA) are debatable[2]; however, there are some findings indicative of an idiosyncratic drug reaction, pointing to a genetic basis of this adverse effect.[3] Some studies suggest that specific HLA haplotypes are associated with a patient's susceptibility to developing CA.[4-8] In these studies, HLA subtyping of Jewish and non-Jewish Caucasian patients with and without CA was performed. Clozapine-induced agranulocytosis was associated with HLA-B38, DRB1*0402, DRB4*0101, DQB1*0201, and DQB1*0302 haplotypes in Jewish and HLA-DR*02, DRB1*1601, DRB5*02, and DQB1*0502 in non-Jewish Caucasian patients[4-7]; the presence of HLA-B35 seemed to be a protective factor against CA in the latter group.[8] However, each of these studies had statistical and/or methological shortcomings, such as small sample size, lack of clinical descriptions of study subjects, or insufficient information about the study design.

In 1996 we initiated a collaborative study with the Drug Commission of the German Medical Association (DCGMA). Within its pharmacovigilance program, the DCGMA receives encoded information about adverse drug reactions reported by clinicians and hospitals nationwide. In cases of reported CA (defined as an absolute neutrophil count of less than 500×10^9/L), a letter was sent to the reporting physician, providing information about our study and requesting a blood sample of his or her patient after written informed consent had been given by the patient. Recently, 30 patients with CA (17 women, 13 men; mean age, 44.5 and 49.8 years, respectively) and 77 age-related patients (40 women, 37 men), who were treated with clozapine for at least 2 years without developing CA, have been included in our study. This makes our sample, together with the study by Claas et al,[8] the largest non-Jewish Caucasian HLA-subtyped CA sample. All subjects of both groups met *DSM-III-R* criteria for schizophrenia (paranoid type), received no other concomitant medication, and were German Caucasians. Moreover, the clozapine dose in both groups was comparable (mean, 225 mg vs 275 mg).

Polymerase chain reaction–based techniques were used to identify specific alleles of HLA-B, DRB, and DQB in patients and controls. Reaction conditions were applied as recommended by the manufacturers. Frequencies were compared by χ^2 test (or by Fisher exact test when a cell frequency was 5 or less. Results are presented in Table 1 as nominal *P* values without adjusting for

INDEX OF FIGURES AND

TABLES

multiple hypothesis tests, since controlling th ov rall type I error was of less concern than inflating type II error[9] with respect to CA.

The trend for a higher frequency of HLA-DQB*0201 in Caucasian patients with CA ($P<.07$) seems of particular interest, since this haplotype has been defined by Amar et al[7] as a genetic marker for Jewish Caucasian patients with CA. This HLA haplotype might represent a common genetic marker for CA in patients of different ethnic groups. However, this assumption needs to be confirmed in a larger cohort of patients, as do our findings of a significant higher frequency of HLA-DQB*0502 and DRB5*02 in non-Jewish Caucasians with CA. Other previously reported HLA antigens of clinical relevance to CA were equally distributed in our CA sample. In conclusion, we could confirm some but not all prior findings of HLA subtyping in Caucasian patients with CA. However, in our opinion, HLA associations with CA represent only 1 of several explanations for the hypothesized genetic background of CA. Likewise, hereditary polymorphisms of specific metabolizing enzyme systems or relevant receptors could be involved in the pathomechanisms of this idiosyncratic drug reaction.

Michael Dettling, MD
Department of Psychiatry
University Clinic Benjamin Franklin
Free University of Berlin
Eschenallee 3
D-14050 Berlin
Germany

Ingolf Cascorbi, MD, PhD
Ivar Roots, MD
Bruno Mueller-Oerlinghausen, MD
Berlin

INDEX OF FIGURES AND TABLES

1. Alvir JM, Lieberman JA, Safferman AZ, Schwimmer JL, Schaaf JA. Clozapine-induced agranulocytosis: incidence and risk factors in the United States. *N Engl J Med.* 1993;329:162-167. MEDLINE

2. Krupp P, Barnes P. Clozapine-associated agranulocytosis: risk and aetiology. *Br J Psychiatry.* 1992;17:38-40.

3. Uetrecht JP. Idiosyncratic drug reactions: possible role of reactive metabolites generated by leukocytes. *Pharmacol Res.* 1989;6:265-273.

4. Lieberman JA, Yunis J, Egea E, Canoso RT, Kane JM, Yunis EJ. HLA-B 38, DR 4, DQw3 and clozapine-induced agranulocytosis in Jewish patients with schizophrenia. *Arch Gen Psychiatry.* 1990;47:945-948. MEDLINE

5. Yunis JJ, Corzo D, Salazar M, Lieberman JA, Howard A, Yunis EJ. HLA-associations in clozapine-induced agranulocytosis. *Blood.* 1995;86:1177-1183. MEDLINE

6. Valevski A, Klein T, Gazit E, Meged S, Stein D, Elizur A, Narinsky ER, Kutzuk D, Weizman A. HLA-B 38 and clozapine-induced

agranulocytosis in Israeli Jewish schizophrenic patients. *Eur J Immunogenet.* 1998;25:11-13.

7. Amar A, Segman R, Shtrussberg S, Sherman L, Safirman C, Lerer B, Brautbar C. An association between clozapine-induc d agranulocytosis in schizophrenics and HLA-DQB1*0201. *Int J Neuropsychophamacol.* 1998;1:41-44.

8. Claas FHJ, Abott PA, Witvliet MD, Amaro JD, Barnes PM, Krupp P. No direct clinical relevance of the human leukocyte antigen (HLA-) system in clozapine-induced agranulocytosis. *Drug Safety.* 1992;7 (suppl 1):3-6.

9. Pernegger TV. What's wrong with Bonferroni adjustments. *BMJ.* 1998;316:1236-1238. MEDLINE

This work was supported by grant 01 EC 9406 (Dr Dettling) from the Federal Ministry of Education and Research, Bonn, Germany.

ARCHIVES
OF
GENERAL PSYCHIATRY

Vol. 58 No. 1,
January 2001

Table

Return to Index of Figures and Tables

Genetic Determinants of Clozapine-Induced Agranulocytosis: Recent Results of HLA Subtyping in a Non-Jewish Caucasian Sample

(*Arch Gen Psychiatry.* 2001;58:93-94)

HLA Subtyping of Schizophrenic Subjects With CA (Cases, N = 30) and Age- and Sex-Matched Subjects With Schizophrenia Without CA (Controls, N = 77)*

	No. (%)			
HLA	Cases	Controls	Odds Ratio (95% CI)	*P*
B35	5 (17.2)	14 (18.2)	0.94 (0.31-2.89)	...
B38	5 (17.2)	9 (11.7)	1.57 (0.48-5.16)	...
DQB1*0201	13 (43.3)	20 (25.6)	2.22 (0.92-5.36)	.07
DQB1*03	16 (53.3)	43 (55.1)	0.93 (0.40-2.17)	...
DQB1*0302	5 (16.7)	12 (15.4)	1.10 (0.35-3.44)	...
DQB1*0502	5 (16.7)	1 (1.3)	15.4 (1.72-138)	.006
DRB1*0402	3 (10.0)	2 (2.6)	4.22 (0.67-26.6)	...
DRB1*1601	3 (10.0)	3 (3.8)	2.78 (0.53-14.6)	...
DRB4	11 (36.7)	35 (44.9)	0.71 (0.30-1.69)	...
DRB5*02	3 (10.0)	0 (0.0)	NA	.02

** CA indicates clozapine-induced agranulocytosis; CI, confidence interval; ellipses, not significant; and NA, not applicable.*

Please close this browser window to return to the article.